Online Figure-Ground Segmentation with Adaptive Metrics in Generalized LVQ

Alexander Denecke[1,2], Heiko Wersing[2], Jochen J. Steil[1], Edgar Körner[2]

1- Bielefeld University - CoR-Lab
P.O.-Box 10 01 31, D-33501 Bielefeld - Germany
adenecke@cor-lab.uni-bielefeld.de
2- Honda Research Institute Europe
Carl-Legien-Str. 30, D-63073 Offenbach/Main - Germany

Abstract

We address the problem of fast figure-ground segmentation of single objects from cluttered backgrounds to improve object learning and recognition. For the segmentation, we use an initial foreground hypothesis to train a classifier for figure and ground on topographically ordered feature maps with Generalized Learning Vector Quantization. We investigate the contribution of several adaptive metrics to enable generalization to the main object parts and derive a foreground classification, which yields an improved bottom-up hypothesis. We show that metrics adaptation is a powerful enrichment, where generalizing the Euclidean metrics towards local matrices of relevance-factors leads to a higher classification accuracy and considerable robustness on partially inconsistent supervised information. Additionally, we verify our results in an online learning scenario and show that figure-ground segregation using this adaptive metrics enables a considerably higher recognition performance on segmented object views.

Key words:
relevance learning, figure-ground segregation, generalized learning vector
quantization, object recognition

1 Introduction

For research in human-machine interaction, the learning of visual representations under general environmental conditions becomes increasingly important. The main goal is to reach a symbolic level for a compact and unambiguous description of the visual data. Therefore the segregation of objects from their

surrounding background is fundamental for object learning and recognition. The problem for segmentation is to group similar parts of the scene to each other. As the notion of *similar* is not clearly defined, this problem can be addressed in several ways and by the usage of different information sources. Possible criteria for similarity are the homogeneity of regions, coherent motion or semantic properties. In the following we will give an overview of current state of the art methods for segregating an object from the surrounding background.

In general, most models represent the image data by a stack of topographically ordered feature-maps (e.g. color, texture and edge detections) with one feature for every pixel position. The problem of figure-ground segregation then reduces to the problem of assigning the corresponding feature representatives to figure or ground. In the following, we will separate the segmentation approaches into three categories: *object-specific models* that use learnt knowledge about particular objects in a top-down fashion, *bottom-up models* that generate a segmentation entirely based on the feature similarities for each new image, and *hypothesis-driven models* that use a prior coarse hypothesis on figure and ground to obtain a precise segmentation of an object.

A prominent example for object-specific models are the parts-based approaches [1,2,3,4], whose goal is to model an object class/category by a set of typical image patches obtained by a learning algorithm. Such a representation can be used to detect corresponding patches in the target images to find/recognize the objects, as well as to segment them from the background. Therefore these methods can be assigned to the class of top-down models. The concept of parts can also be generalized to more complex structures [5]. The general problems of these methods are the high computational load in the learning phase, as well as the necessity of a database to acquire the representation. For interactive scenarios where real-time and online processing are significant constraints these models are currently not appropriate.

The bottom-up segmentation models avoid referencing to a particular object specific representation. With the Normalized Cuts Method [6] the whole image is modeled by an interaction matrix, representing all pairwise feature similarities. The goal is to partition a graph defined by the interaction matrix into two regions with strong self-similarities but only weak connections to the other region. The Competitive Layer Model has been designed as a dynamic model of Gestalt-based feature binding and segmentation [7] using similar pairwise feature similarities. The data-driven learning of these similarity functions has been considered by Weng et al. [8]. But such approaches solve complex optimization problems resulting in computationally demanding models, which are also not appropriate for online learning.

Hypothesis-driven approaches model the feature distribution of figure and

ground and combine them with constraints on the derived foreground regions. Additionally, e.g. the similarity information of neighboring pixels can be used to derive consistent segments respecting the homogeneities and discontinuities in the image. For example, Rother et al. [9] propose to model foreground and background by Gaussian Mixture Models (GMM) and use the Min-Cut algorithm to optimize the partition of the image into two regions with respect to the model affinity and discontinuities in the image. As the basic Grab-Cut [9] model is sensitive to high contrast edges in cluttered background, Sun et al. [10] suppress this effect with information from the known and static background. Similar to GMM, Weiler et al. [11] uses histograms for the region description integrated into a Level-Set energy functional with an included smoothness term (e.g. penalizing the length of the contour) to derive compact foreground segmentations. The methods of Rother et al. and Weiler et al. [9,11] rely on the necessity of sufficient image statistics to model the feature distributions and high color gradients across figure-ground boundaries to align the segmentation with the object contour. In [12], the clusters in the color-space of the image are modeled with prototypical feature combinations. This concept is generalized to arbitrary feature-maps, for example to derive compact regions in the image space by a direct integration of the pixel position as additional features [13]. The latter two approaches [13,12] select the supposed foreground clusters to derive a segmentation. For this selection the concept of a segmentation hypothesis is needed.

The hypothesis-driven methods do not need an object specific training beforehand, but an initial guess which parts of the image are related to figure and ground to obtain the segmentation. This initial guess can be derived from foreground detection [10], user interaction [9], depth information [13] or saliency [12]. It is a common problem that the obtained hypothesis has a noisy character, caused by fundamental problems (e.g. the ill-posed task of depth estimation from 2D data). Therefore the main problem is to generalize to relevant object regions from such imprecise hypotheses. One approach is to obtain a model classifying foreground and background based on the pixel-wise feature information from the hypothesis. An appropriate learning model can then generalize over inconsistent training data and yields a segmentation that is better than the initial guess (that is, refines the hypothesis). This concept can be transferred to other application domains as well, like audio segmentation.

The segmentation obtained by such hypothesis-driven models can be combined with a high level object representation used for learning and recognition of the segmented object views. In the context of online learning of objects a biologically inspired view-based approach on the basis of hierarchically organized processing was proposed recently [14]. Using this model as part of an active stereo vision system (see Sec. 2), object learning and recognition takes place on the highest level of multiple layers from simple to complex feature detectors (Fig. 1). That is, during the interaction with the user, this method is capable
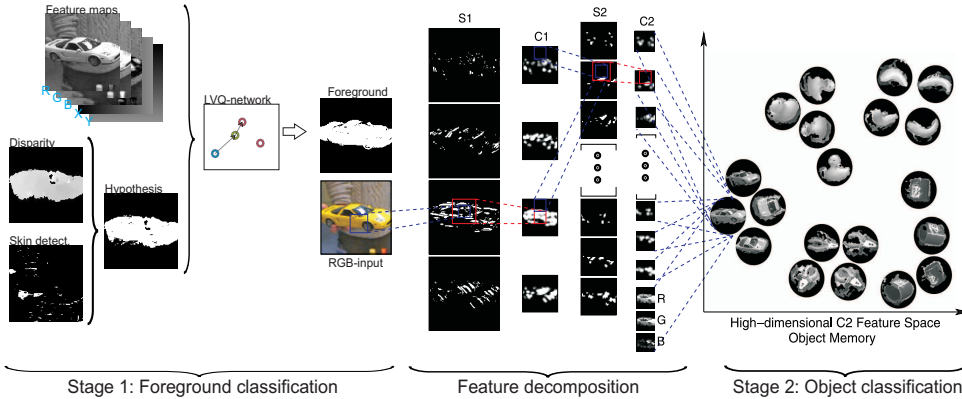
Figure 1. Overview on the architecture for object learning and recognition [14]. In the first stage, color and position are used as features together with the initial hypothesis to obtain the object segmentation with a Learning Vector Quantization approach. In the second stage, the rightmost layer of a feature hierarchy is used as object representation for learning and recognition.

of learning the object representation on the basis of high dimensional shape features. For this architecture it was shown that the performance of the object classifier improves considerably with better segregation from the background [13].

Following the general architecture shown in Fig. 1, we propose a hypothesis-driven method to segment objects for object learning that is capable of running with sufficient speed and can handle changing and cluttered backgrounds. We assume that an initial hypothesis from depth estimation is given, which covers the image region of the presented object. Then our method for object segmentation uses prototypical feature representatives to model figure and ground. Because extracting 3D information from 2D images in general is an ill-posed problem, the resulting hypothesis is characterized by a partially inconsistent overlap with the outline/region of the object (see Fig. 2). We use this information as a supervised label for the image features to train a classifier for figure and ground with Generalized Learning Vector Quantization (GLVQ [15]). The goal is to generalize to the main object parts and to derive a foreground classification, which improves the initial hypothesis. In prototype-based representations, the clustering and classification of image regions on the basis of similarity crucially depends on the underlying metrics. For GLVQ several extensions of the Euclidean metrics are available [16,17], which offer additional feature and prototype-specific weighting factors, taking into account the discriminative power of features and correlations between them. These so-called relevance-factors and the LVQ-network weights (prototypes) are adapted online by means of gradient descent. By comparing the adaptive metrics and investigating the robustness to the noisy supervised information, we show that manipulating the metrics given a prototypical feature representation is capable of achieving a large gain in hypothesis refinement. Transferring these insights

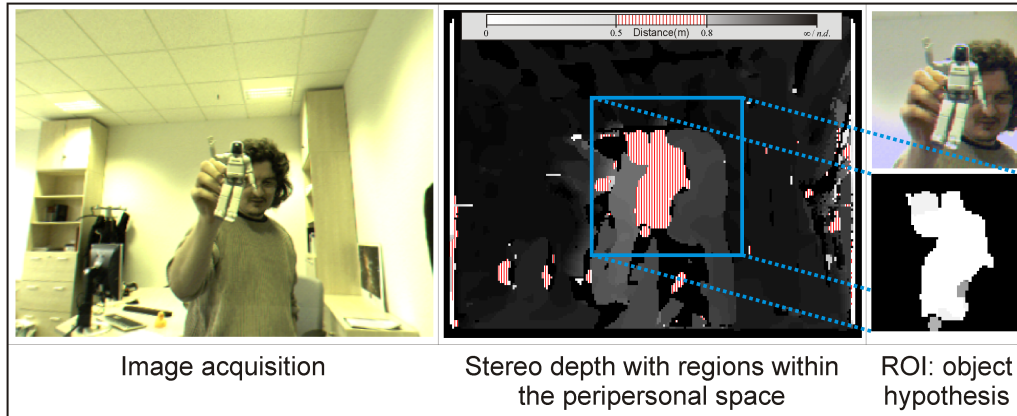| Image acquisition | Stereo depth with regions within the peripersonal space | ROI: object hypothesis |

Figure 2. Overview on the scenario for object learning and recognition. To determine an initial hypothesis that defines which parts of the view correspond to the object itself, the motion and depth information is used for attending and selecting the object during interaction. For this, the concept of peripersonal space [14] is used, which defines the behaviorally relevant parts of the visual scene as the region in front of the system. The highlighted region in the middle image consists of all scene elements within a specified depth interval.

to the application domain of figure-ground segregation, we show that the introduction of prototype-specific matrices of relevance-factors is leading to an improved segmentation quality enhancing object learning and recognition. In contrast to other prototype-based approaches [13,12], this method offers the advantage to automatically determine those feature dimensions most relevant for the object segmentation. Additionally, it relaxes a priori assumptions on object position and segment selection.

The paper is organized as follows. First we present our current scenario and concept for object segmentation. After a short description of four adaptive metrics extensions for GLVQ, we compare them with respect to foreground classification performance on multidimensional feature vectors, in particular with noisy training data. Finally we evaluate the impact of metrics adaptation on the prototype-based representations on real world data for object recognition and compare the proposed segmentation method to another algorithm on a benchmark dataset.

2 Method

Our current scenario for object learning consists of a user presenting objects to a stereo-camera system. For unconstrained interaction, the pan-tilt stereo-camera head is controlled by an attention system for object localization and tracking. The behaviorally relevant parts of the scene for learning and recognition are defined by the concept of peripersonal space (Fig. 2). According

to this concept, the depth estimation of the region in front of the system is analyzed with a blob-detection within a specified depth interval (in this work 50cm-80cm). The most salient blob is tracked by the system and centered in view by setting the gaze direction. This assures invariance to the location of the object in the scene (translation invariance). From the blob-detection a square region of interest (ROI) is defined based on a distance estimate and normalized to a size of $\mathcal{I} \times \mathcal{I}$ pixels, where we use $\mathcal{I} = 144$. This assures, that the same object which is presented in different distances to the learning systems is processed with nearly the same size (size invariance). To improve the learning and recognition of the objects, the goal is to segment the object from the RGB image of the ROI while the depth-information (respectively the binarization) is used as initial hypothesis $\mathcal{H}$ (upper right and lower right image of Fig. 2). Object learning is then based on the segmented object views (compare Fig. 1).

2.1 Problem description

Extracting 3D information from 2D images in general is an ill-posed problem and results in coarse approximations of the object outline/region by the depth estimation. Therefore, generalizing to the relevant object parts from this hypothesis and discarding the background is complicated. This is caused by partially overlapping feature-clusters due to the noisy hypothesis, as well as by similar colors in regions of figure and ground. Formally, the input data consist of a stack of $M = 5$ feature maps $\mathcal{F} := \{F_i | i = 1..M\}$ corresponding to the RGB color-space and pixel position ($F_1^{x,y} = R^{x,y}$, $F_2^{x,y} = G^{x,y}$, $F_3^{x,y} = B^{x,y}$, $F_4^{x,y} = x$, $F_5^{x,y} = y$). The choice of these features is not constrained to a particular color space and other features like texture could be included as well. The pixel coordinates (x, y) are important as additional features for an implicit region modeling. The stack of maps $\mathcal{F}$ is represented by a set of vectors $\vec{\xi} \in \mathbb{R}^M$, where every pixel defines a feature vector $\vec{\xi}^{x,y} = (F_1^{x,y}..F_M^{x,y})^T, 1 \leq x, y \leq \mathcal{I}$. We assume an unknown ground truth map $\mathcal{G}$, which defines the membership of feature $\vec{\xi}^{x,y}$ for every pixel (x, y) to figure $\mathcal{G}^{x,y} = 1$ or ground $\mathcal{G}^{x,y} = 0$ with respect to the attended object. The goal is to approximate $\mathcal{G}$ by a binary map $\mathcal{A}$ using the initial hypothesis $\mathcal{H}$ (also a binary map) and the similarity information provided from the feature-maps $\mathcal{F}$. The binary foreground map $\mathcal{A}$ is the result of a pixel-wise foreground classifier, which is trained on features $\mathcal{F}$ and hypothesis $\mathcal{H}$ for the current image, $\mathcal{A}^{x,y} \leftarrow \mathcal{A}_{\mathcal{F},\mathcal{H}}^{x,y}(\xi^{x,y})$. Though we cannot expect that the ground truth map can be fully recovered by $\mathcal{A}$, the goal is to discard at least the inconsistent parts of the hypothesis. If the ground truth information is available, the segmentation quality can be quantified by a pixel-wise comparison of $\mathcal{G}$ with the resulting foreground classification $\mathcal{A}$, i.e. $D(M_1 = \mathcal{A}, M_2 = \mathcal{G}) := 1 - \dfrac{\sum_{x,y} |M_1^{x,y} - M_2^{x,y}|}{\mathcal{I}^2}$. But using this pixel-wise

comparison, one must be aware of the variability of the foreground hypotheses/segmentation in their size and proportion to the number of background pixels within the sequence of images. Therefore we measure the success of the segmentation by an increased overlap $S(\mathcal{A}, \mathcal{G}) > S(\mathcal{H}, \mathcal{G})$ of $\mathcal{A}$ with the *ground truth* segmentation $\mathcal{G}$. The similarity function $S(M_1, M_2)$ normalizes the difference of two binary maps $M_1^{x,y}, M_2^{x,y} \in \{0, 1\}$ by the sum of their foreground regions and discards the background pixels.

$$S(M_1, M_2) := 1 - \frac{\sum_{x,y} |M_1^{x,y} - M_2^{x,y}|}{\sum_{x,y} M_1^{x,y} + \sum_{x,y} M_2^{x,y}}$$

This measure $S(M_1, M_2)$ yields a monotonically increasing function dependent on the overlap of M_1 and M_2. Note that, if the figure occupies only a small fraction of the image, then $S(\mathcal{A}, \mathcal{G})$ and $D(\mathcal{A}, \mathcal{G})$ can be strongly different, because the latter is mainly computed on the background.

2.2   *General concept for segmentation*

After the acquisition of the feature maps $\mathcal{F}$ and the hypothesis $\mathcal{H}$ (Fig. 2) a pre-processing $F_i^{x,y} \leftarrow T_F(F_i^{x,y})$ of the feature maps $F_i^{x,y}$ (a gamma correction and white balancing on the maps representing the image data) is performed first. Afterwards all skin-colored areas $\mathcal{S}, \mathcal{S}^{x,y} \in \{0, 1\}$ (filtered in a separate processing stream for skin color detection [18]) are removed ($T_H(\mathcal{H}) := \mathcal{H} \leftarrow \mathcal{H} - (\mathcal{H} \cap \mathcal{S})$) from the hypothesis $\mathcal{H}$. This is necessary because the hand is strongly connected to every object/hypothesis and state of the art object classifiers are not capable of learning/representing the special role of the skin colored areas. To build $\mathcal{A}$ and to extract the relevant object parts from $\mathcal{F}$ using $\mathcal{H}$, we state the task of object segmentation as a binary classification problem and use generalized learning (i.e. supervised) vector quantization to train a classifier for foreground. We adapt a codebook of N class-specific prototypes $\mathcal{P} := \left\{ \vec{w}_p \in \mathbb{R}^M | p = 1..N \right\}$, to represent the clusters in the data $\mathcal{F}$ (homogeneous regions in the image) by the prototypes $\vec{w}_p$. For figure-ground segregation a setup with two classes is used where $c(\vec{w}_p) \in \{0, 1\}$ encodes the class-membership, assigned by the user, of every prototype to figure or ground. The codebook $\mathcal{P}$ is initialized for each class separately with a random sampling of features $\vec{\xi}$ from the first image (respectively $\mathcal{F}, \mathcal{H}$). After the initialization of $\mathcal{P}$, this codebook is adapted for every succeeding image (Sec. 2.3) on randomly chosen pairs $(\vec{\xi}^{x,y}, \mathcal{H}^{x,y})$. The reuse of prototypes on subsequent images accounts for the continuity of the image sequence and allows a reduced number of update steps on a single image. In the evaluation-phase, the image is partitioned into N segments (binary maps) $V_p \in \{0, 1\}$ by assigning all feature vectors $\vec{\xi}^{x,y}$ (i.e. pixels) independently to the prototype $\vec{w}_p$ with the smallest distance $d(\vec{\xi}^{x,y}, \vec{w}_p)$. Using an adaptive

Learning Vector Quantization approach, the final segmentation $\mathcal{A}$ is combined by choosing the activation-maps from prototypes assigned to the foreground $\mathcal{A} = \sum_p^N c(\vec{w}_p) V_p$.

The general concept for combining the information from the image and the hypothesis can be summarized with the following pseudo code:

(1) Input: feature maps and hypothesis from object ROI:
$\mathcal{F}^{x,y} := \{F_i^{x,y} | i = 1..M\}$,
$\mathcal{H}^{x,y} \in \{0, 1\}$

(2) Preprocessing of feature maps:
$F_i^{x,y} \leftarrow T_F(F_i^{x,y})$

(3) Preprocessing of hypothesis:
$\mathcal{H} \leftarrow T_H(\mathcal{H})$

(4) Init codebook $\mathcal{P} = \{\vec{w}_p\}, p = 1, .., N$ if not already done

(5) Adaptation (for t update steps)
- Select $\vec{\xi}^{x,y}$ at random position $1 \le x, y \le \mathcal{I}$
- Find best matching prototypes $\vec{w}_J$ for the correct label, $\vec{w}_K$ for the incorrect label
$$\vec{w}_J = \{\vec{w}_p \in \mathcal{P} | d(\vec{w}_p, \vec{\xi}^{x,y}) = \min_{q, c(\vec{w}_q) = \mathcal{H}^{x,y}} d(\vec{w}_q, \vec{\xi}^{x,y})\}$$
$$\vec{w}_K = \{\vec{w}_p \in \mathcal{P} | d(\vec{w}_p, \vec{\xi}^{x,y}) = \min_{q, c(\vec{w}_q)!=\mathcal{H}^{x,y}} d(\vec{w}_q, \vec{\xi}^{x,y})\}$$

- Update prototypes with learning rate α
$\vec{w}_J \leftarrow \vec{w}_J + \alpha \cdot \Delta \vec{w}_J$
$\vec{w}_K \leftarrow \vec{w}_K + \alpha \cdot \Delta \vec{w}_K$

- Update metrics $d(\cdot, \cdot)$, see Sec. 2.3

(6) Evaluation: for all pixels $1 \le x, y \le \mathcal{I}$
- Compute activation map for each prototype
$$V_p^{x,y} := \begin{cases} 1 & \text{if } d(\vec{\xi}^{x,y}, \vec{w}_p) < d(\vec{\xi}^{x,y}, \vec{w}_r), \forall r \ne p, \{r, p\} \in \mathcal{P}, \\ 0 & \text{else} \end{cases}$$
- Determine foreground segmentation
$\mathcal{A} = \sum_p^N c(\vec{w}_p) V_p$

The ASDF model [13], which is used for the comparison of the performance in Sec. 4, differs in three aspects. In their more heuristical setting, Steil et al. considered an unsupervised clustering approach and therefore only w_J is adapted in step (5) where $c(\vec{w}_p) = 1, \forall p \in \mathcal{P}$, is equal for all prototypes. After adapting the prototypes, the foreground segmentation (6) is constructed with a heuristics to determine a subset of V_p, each of which shows a sufficient overlap with the initial hypothesis $\mathcal{H}$. Additionally to the original hypothesis derived from depth and skin color information a further position prior, an image centered circular map is used. The most important difference concerns the distance computation, which is Euclidean and not adapted during learning.

Similarity-based clustering and classification crucially depends on the underlying metrics and many modifications of the Euclidean metrics have been proposed. One of the most popular metrics manipulation is the introduction of feature-specific weighting factors, for example to compensate for different scales of the feature channels. The ASDF approach globally modifies the metrics by a rescaling of the feature maps $T_F(F_i^{x,y}) := f_i \cdot (F_i^{x,y}/\sigma_i^2)$ with their variance σ_i^2 and a feature-specific a priori weighting factor f_i. However, finding the appropriate weightings is a tough problem. Recently, for Learning Vector Quantization it has been proposed to optimize such factors for the classification problem at hand. Based on the Generalized LVQ (GLVQ [15]) method, Hammer [16] has extended the standard Euclidean metrics by introducing a global relevance-factor for each feature dimension (Generalized *Relevance* LVQ (GRLVQ)). This leads to the squared weighted Euclidean metrics

$$d(\vec{\xi}, \vec{w}) = \|\vec{\xi} - \vec{w}\|_\lambda^2 = \sum_i^M \lambda_i(\xi_i - w_i)^2,$$

where $\lambda_i \geq 0$ and $\sum_{i=1}^M \lambda_i = 1$. In further investigations, the following two extensions of this concept have been proposed [19]. First, using an $M \times M$ matrix of relevance-factors (Generalized Matrix LVQ, GMLVQ results in the metrics

$$d(\vec{\xi}, \vec{w}) = (\vec{\xi} - \vec{w})^T \Lambda (\vec{\xi} - \vec{w}),$$

where Λ is positive semi-definite, assured by adapting Ω, where $\Lambda = \Omega\Omega^T$ to yield a valid metrics, i.e. $d(\vec{\xi}, \vec{w}) = (\vec{\xi} - \vec{w}_p)^T \Omega\Omega^T (\vec{\xi} - \vec{w}_p) = (\Omega^T(\vec{\xi} - \vec{w}_p))^2 \geq 0$. Additionally, the authors advise to normalize the diagonal elements by $\sum_{i=1}^M \Lambda_{i,i} = 1$ to stabilize the algorithm. The second extension introduces local relevance-vectors/matrices $\vec{\lambda}_p, \Lambda_p$ specific for every prototype, called localized GMLVQ/GRLVQ (LGMLVQ/LGRLVQ) to allow prototype specific metrics manipulations, i.e. $d(\vec{\xi}, \vec{w}_p) = (\vec{\xi} - \vec{w}_p)^T \Lambda_p (\vec{\xi} - \vec{w}_p)$. As introduced by GLVQ, the overall performance of the network is measured by

$$E = \sum_{\vec{\xi}^{x,y}} \sigma\left(\mu(d)\right),$$

$$\sigma(x) = \frac{1}{1+e^{-x}}, \ \mu(d) = \frac{d_J - d_K}{d_J + d_K},$$

$$d_J = d(\vec{\xi}^{x,y}, \vec{w}_J), \ d_K = d(\vec{\xi}^{x,y}, \vec{w}_K).$$

The error E is minimized on training samples $(\vec{\xi}^{x,y}, \mathcal{H}^{x,y})$, where d_J is the distance between $\vec{\xi}^{x,y}$ and the most similar prototype from the correct class with $\mathcal{H}^{x,y} = c(\vec{w}_J)$ and d_K is the distance to the most similar prototype from an incorrect class. Using stochastic gradient descent to minimize E, the

prototypes $\vec{w}_p$ of the network and the relevance-factors $\vec{\lambda}, \Lambda$ are updated by $\vec{w} \leftarrow \vec{w} + \alpha \cdot \Delta\vec{w}$, $\vec{\lambda} \leftarrow \vec{\lambda} + \beta \cdot \Delta\vec{\lambda}$. See [20] for a comprehensive overview and the derivations of the update formulas. For the most complex case, LGMLVQ, the prototypes as well as the relevance matrices of the two nearest prototypes $\vec{w}_J$ and $\vec{w}_K$ are adapted by means of:

$$
\begin{aligned}
\Delta\vec{w}_J &= \frac{\partial E}{\partial \vec{w}_J} = \frac{\partial \sigma}{\partial \mu}\frac{\partial \mu}{\partial d_J}\frac{\partial d_J}{\partial \vec{w}_J} = -\alpha \cdot \frac{e^{-\mu}}{(1+e^{-\mu})^2}\frac{2d_K}{(d_J+d_K)^2}(-2\Omega\Omega^T(\vec{\xi}-\vec{w})), \\
\Delta\vec{w}_K &= \frac{\partial E}{\partial \vec{w}_K} = \frac{\partial \sigma}{\partial \mu}\frac{\partial \mu}{\partial d_K}\frac{\partial d_K}{\partial \vec{w}_K} = \alpha \cdot \frac{e^{-\mu}}{(1+e^{-\mu})^2}\frac{2d_J}{(d_J+d_K)^2}(-2\Omega\Omega^T(\vec{\xi}-\vec{w})), \\
\Delta\Lambda_J &= \frac{\partial E}{\partial \Lambda_J} = \frac{\partial \sigma}{\partial \mu}\frac{\partial \mu}{\partial d_J}\frac{\partial d_J}{\partial \Lambda_J} = -\beta \cdot \frac{e^{-\mu}}{(1+e^{-\mu})^2}\frac{2d_K}{(d_J+d_K)^2} \cdot (M_J^T + M_J), \\
\Delta\Lambda_K &= \frac{\partial E}{\partial \Lambda_K} = \frac{\partial \sigma}{\partial \mu}\frac{\partial \mu}{\partial d_K}\frac{\partial d_K}{\partial \Lambda_K} = \beta \cdot \frac{e^{-\mu}}{(1+e^{-\mu})^2}\frac{2d_J}{(d_J+d_K)^2} \cdot (M_K^T + M_K), \\
M_J &= \Omega(\vec{\xi}-\vec{w_J}) \cdot (\vec{\xi}-\vec{w_J})^T, \\
M_K &= \Omega(\vec{\xi}-\vec{w_K}) \cdot (\vec{\xi}-\vec{w_K})^T.
\end{aligned}
$$

To keep a compact notation, in the following we will refer to the Generalized Vector Quantization with the symbol $\mathcal{Q}$ and use the indices L, G for localized or global metrics extension and M, V for the relevance matrices Λ or vectors $\vec{\lambda}$. That is, GLVQ=$\mathcal{Q}$, GRLVQ=$\mathcal{Q}_V^G$, GMLVQ=$\mathcal{Q}_M^G$, LGRLVQ=$\mathcal{Q}_V^L$, LGMLVQ=$\mathcal{Q}_M^L$.

The relevance factors of $\mathcal{Q}_V^G/\mathcal{Q}_V^L$ yield an ellipsoidal-shaped, axis-parallel scaling of data points equidistant to a prototype. In the case of the matrix transformations the distance computation is shaped to a rotated ellipsoidal. In the simplest case of only one prototype for each class, standard GLVQ with the Euclidean metrics separates two classes by a linear hyperplane (the border of the Voronoi cells). This behavior does not change with the introduction of global transformations ($\mathcal{Q}_M^G, \mathcal{Q}_V^G$). On the contrary, the extension of local relevance transformations introduces more flexible (non-linear) decision boundaries between each pair of prototypes, by using different metrics for them. This effect is independent of the usage of multiple prototypes which yields more complex tessellations of the feature space. The adaptive metrics are of special interest for our scenario due to the capability to weight the features according to their relevance for the classification task. The main idea of the matrix transformation is to account for correlations/combinations of the feature dimensions in the off-diagonal elements of Λ.
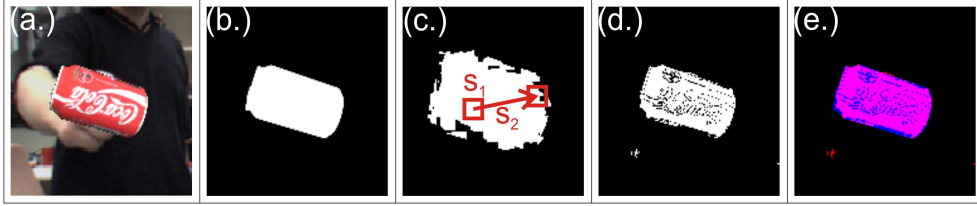
Figure 3. Example image from the dataset of rendered objects and corresponding distortion of the ground truth data. From left to right the original image (a), ground truth $\mathcal{G}$ (b), distorted hypothesis $\mathcal{H}$ (c) with a patchsize $s_1 = 12$, shift $s_2 = 22$ and the resulting segmentation $\mathcal{A}$ (d) derived by a classifier trained on $\mathcal{H}$. Finally the visualization of the overlap of (b) and (d), which is quantified by the measure $S(\mathcal{A}, \mathcal{G})$ during the experiments.

3 Investigation of metrics adaptation

We have formulated the segmentation such that a noisy hypothesis is used to train a classifier for figure and ground using the samples $(\vec{\xi}^{x,y}, \mathcal{H}^{x,y})$. From previous results in [17] it is known that the performance in classification benchmarks can strongly benefit from the usage of the adaptive metrics. According to our hypothesis-driven learning approach, we investigate the impact of these methods with respect to the quality of the target information $\mathcal{H}$ and the generalization capabilities to the relevant image structures. After the description of the general setup used for our experiments, we consider the results from a single image in a simple example to get some insights what happens with increasing noise. Secondly we use ground truth data from a rendered-object dataset to compare the different adaptive metrics by their capability of optimizing the classifier on the basis of an existing set of prototypes. Thirdly we investigate the generalization capabilities of LGMLVQ by using different levels of noisy hypotheses, and compare the obtained foreground classification to the ground truth.

3.1 Setup

3.1.1 Database of rendered objects

To investigate the effect of different adaptive metrics in GLVQ we employ a dataset of rendered objects according to our scenario. A collection of rendered image sequences from 25 realistic 3D objects (bottles, boxes, cars etc.) is used, where a ground truth segmentation is available for every object view. The arbitrarily rotated object-views are pasted in the center of a typical non-rendered scene (human in the background, hand near object, see Fig. 3a), generated by tracking the view-centered hand in front of the camera system. Additionally, the corresponding ground truth membership $\mathcal{G}$ of pixels to the foreground is

11

used to generate artificial (noisy) hypothesis maps $\mathcal{H}$ (Fig. 3c). The distortion mimics the noise obtained from standard stereo depth algorithms. This is achieved by randomly selecting and shifting 1000 patches with size $s_1 \times s_1$ from one position in the mask $\mathcal{G}$ to another by a distance randomly chosen between 1 and s_2. To address the capability of hypothesis refinement on the feature-maps $\mathcal{F}$, these hypotheses $\mathcal{H}$ are used as target labels for the randomly chosen pixels during the adaptation of the classifier. During the experiments we generate hypothesis maps with increasing noise by setting $s_1 = 30$ and varying the parameter s_2. The intensity of the scrambling and the similarity of the produced foreground classifications $\mathcal{A}$ to the ground truth data $\mathcal{G}$ and hypothesis $\mathcal{H}$ are quantified by $S(\mathcal{H}, \mathcal{G})$, $S(\mathcal{A}, \mathcal{G})$, $S(\mathcal{A}, \mathcal{H})$, as defined in Sec. 2.1. Due to copyright restrictions on the 3D-objects used for image rendering, the dataset cannot be published. Detailed statistics of the dataset on a per object level are available on request. To give a short overview of the dataset, the average RGB color is (92, 85, 79) for the foreground and (86, 82, 80) for the background. The standard deviation in all feature channels is approximately 55 on both regions. On average the foreground object occupies 13% of the image region, the average bounding box of the images occupies 24%.

The images of the dataset are processed by the method described in Sec. 2.2. For the experiments we use two different configurations for the number of prototypes and learning rates to adapt the networks.

3.1.2   *Multi-prototype setup*

This setup is our current configuration optimized for $\mathcal{Q}_M^L$ to segment the object from the background and is used in our experiments on the complete rendered and realistic datasets. Because we want to investigate the effect of the increasing complexity of the metrics, we use this configuration for all algorithms to ensure comparable conditions. In this configuration, the network consists of $N=20$ randomly initialized prototypes (5 for figure, 15 for ground). The decision on the number of prototypes for both classes depends on the image size, proportion of object size to the background and complexity of foreground and background. Most of the objects presented to the system consist of 3-5 different colors, which explains the choice of 5 prototypes for the foreground class. Note that this does not exclude single colored objects from the segmentation. Typically the background is more complex and cluttered than the foreground such that 10-15 prototypes are appropriate. This decision is supported by observations of Sun et al. [10] and previous experiments with the unsupervised Instantaneous Topological Map (ITM) [21], which was used to estimate the number of prototypes on comparable image data [22].

In particular, we address the figure ground segregation in an online learning scenario. This restricts the computation time to segment each image and in-

troduces constraints on the number of training steps and the learning rates. The prototypes are adapted by 10000 training-steps for each image with a learning rate appropriate for fast adaptation to the changing image content. During preparatory experiments we observed that a fast adaptation of both, prototypes and relevance factors, strongly impairs the performance. By regular sampling in the parameter space spanned by the learning rates, we optimized the learning rates for $\mathcal{Q}_M^L$ towards $\alpha = 0.05$ for the prototype adaptation and $\beta = 0.005$ for the adaptation of the relevance factors. In this setup, to average the prototypes and matrices are effectively updated with values of magnitude around 10^{-4}. While this is moderate for the relevance factors, the prototypes with a range of $\xi_i \in [0..255]$ in the color components are slowly adapted, which is still reasonable on the large amount of data we use (300-700 images per object). Therefore we mainly use metrics learning which is discussed in the experiments. To find an appropriate learning rate for GLVQ, to compare the effect of prototype adaptation and metrics adaptation, also regular sampling in the parameter space was used and yields $\alpha = 100$ for the input data we use in our experiments. Due to the dependence of the effective learning rates on the distances occurring to the best matching prototypes (see $\dfrac{\partial \mu}{\partial d_K}$ in the update rules described in Sec. 2.3), the average update values have a magnitude around 10^{-2}.

3.1.3 Two-prototype setup

For a simple example we use a slightly different setup. First we want to achieve a better separation between the effects of prototype and metrics adaptation and use $\alpha = 0$ to adapt only the metrics. Second we constrain our investigation on a single image and a two class setup, each class modeled by a single prototype $\vec{w}_{fg}, \vec{w}_{bg}$ for foreground and background. This offers the possibility to observe the properties of the prototype under changing noise-conditions and we do not need to account for interactions of multiple prototypes for each class.

3.2 Effect of increasing noise

In this section, we investigate the effect of increasing noise in $\mathcal{H}$ on the data used for training and on the relevance determination of the localized adaptive metrics $\mathcal{Q}_M^L$, $\mathcal{Q}_V^L$. We restrict the experiment to processing a single image, use the two-prototype setup and select an appropriate sample from the dataset of rendered objects consisting of two nearly homogenous regions (Fig. 3a).

In Fig. 4, the corresponding relevance factors for the foreground prototype $\Lambda_{fg}, \vec{\lambda}_{fg}$ as determined by $\mathcal{Q}_M^L$ (left plot) und $\mathcal{Q}_V^L$ (middle plot) are displayed
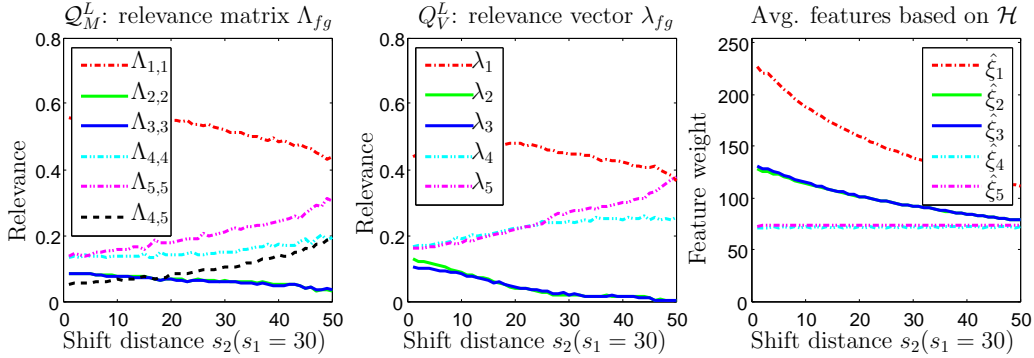
Figure 4. Effect of noise on metrics adaptation of $\mathcal{Q}_M^L$, $\mathcal{Q}_V^L$ on a single image with increasingly distorted hypothesis $\mathcal{H}$ (avg. over 25 repetitions). As the two-prototype setup is used, the prototypes are randomly initialized and not adapted. For $\mathcal{Q}_M^L$ the determined relevance values for the diagonal element of Λ_{fg} corresponding to the color and position as well as the interaction of the pixel position indicated by the off-diagonal element $\Lambda_{4,5}$ are shown. For $\mathcal{Q}_V^L$ the plot contains the components of the relevance vector $\vec{\lambda}_{fg}$. With increased scrambling more and more background is included and changes the properties of the region covered by the hypothesis (right plot). This is indicated by the average $\hat{\xi}_i$ of the feature components in this region. In this case, $\mathcal{Q}_M^L$ and $\mathcal{Q}_V^L$ are capable of adapting the relevance and increase the importance of the coordinates and their interaction.

depending on the increasing noise. For the generation of the plots, the hypothesis was disturbed by 50 levels of noise with fixed window size $s_1 = 30$ and gradually increasing shift distance s_2. To keep conditions on all 50 noise-levels constant, only on the first hypothesis (in this case $\mathcal{H} = \mathcal{G}$) the prototypes have been randomly initialized. This initial set is stored and used for the initialization of the network for the other 50 noise levels. For visualization, the averages of 25 repetitions with different initializations were computed.

With this increasing noise, the properties of the foreground region are continuously changing as observable by the average color features $(\hat{\xi}_1, \hat{\xi}_2, \hat{\xi}_3)$ in the right plot of Fig. 4. As the noise especially affects the object contour, the objects center of mass $(\hat{\xi}_4, \hat{\xi}_5)$ does not change significantly. The average color/position features are computed by $\hat{\xi}_i := \frac{1}{\sum_{x,y} \mathcal{H}^{x,y}} \sum_{x,y} \mathcal{H}^{x,y} \cdot \xi_i^{x,y}$.

Despite the limitations in the setup, the effect of increasing noise on determining the relevance factors can be visualized. The prototypes are not adapted during this experiment ($\alpha = 0$) and therefore not shown. Regarding the relevance factors, the advantage of metrics adaptation becomes visible with an increasingly imprecise hypotheses. That is, the color features become less important than the position, indicated by the changes in their determined relevance. While the center of mass does not change with increasing noise (see Fig. 4 right) for $\mathcal{Q}_V^L$ the weight of this feature dimensions is simultaneously increased. For $\mathcal{Q}_M^L$ this dependence can also be expressed by the corresponding off-diagonal element $\Lambda_{4,5}$. Hence with increasing noise the introduction of

Method	$\mathcal{Q}$	$\mathcal{Q}_V^G$	$\mathcal{Q}_M^G$	$\mathcal{Q}_V^L$	$\mathcal{Q}_M^L$
$S(\mathcal{A}, \mathcal{G})$	0.076	0.423	0.461	0.646	0.926

Table 1
Evaluation on the rendered-object dataset with the multi-prototype setup (i.e. $\alpha = 0.05$). In this table the average similarity of foreground classification $\mathcal{A}$ to ground truth $\mathcal{G}$ for $\mathcal{Q}$ with different adaptive metrics is shown (5 repetitions on 25 objects and 700 views of the dataset). Here the perfect training data $\mathcal{H} = \mathcal{G}$ was used to adapt the classifier. For this $S(\mathcal{A}, \mathcal{G})$ allows conclusions to the foreground classification error introduced by the methods itself. Also we can observe from these results the increase in foreground classification performance caused by the increasing complex metrics adaptation.

the position gets more important for the foreground classification which is the desired behavior. The effect of metrics adaptation compared to the prototype learning will be further evaluated in Sec. 4.

3.3 Learning on ground truth

Here we investigate the capabilities of the adaptive metrics to optimize the classifier on the basis of an existing set of prototypes, i.e. adapt primarily the metrics and only slightly the prototypes. Contrary to the previous experiment, we use the multi-prototype setup with the learning rate $\alpha = 0.05$. Due to the changing image statistics within the large dataset caused by changing background and different objects, a learning rate $\alpha = 0$ is not reasonable. This enables a high flexibility in the metrics adaptation, which yields the best performance in our scenario (Sec. 2.3), and can be regarded as a compromise between plasticity and stability for online learning. For this baseline test (Tab. 1), we apply the variants of GLVQ using different adaptive metrics to the complete dataset of rendered objects. In this experiment we use the ground truth data $\mathcal{H} = \mathcal{G}$ for supervised learning and the complexity of the adaptive metrics is the only modified condition. From Tab. 1 it is visible that an increasing complexity of the adaptive metrics from relevance-vectors to matrices and from global to local ones clearly leads to an improved foreground classification performance and increasing capability to compensate the strongly reduced prototype adaptation. Measured by the overlap S, which considers only foreground-pixels, the resulting foreground mask reaches an average similarity to the ground truth data up to 0.92 for $\mathcal{Q}_M^L$. In particular the results on the whole dataset give a more differentiated view on the capabilities of the different adaptive metrics. While $\mathcal{Q}_M^L$ yields a tolerable testing error (derived from the similarity $S(\mathcal{A}, \mathcal{G})$), the less complex metrics adaptations are not appropriate for an application on the intended scenario. Note that, although $S(\mathcal{A}, \mathcal{G})$ can be very small for $\mathcal{Q}$, the overall pixel-wise classification performance is much better (defined by $D(\mathcal{A}, \mathcal{G})$ in Sec. 2.1), e.g., 87% for $\mathcal{Q}$ and
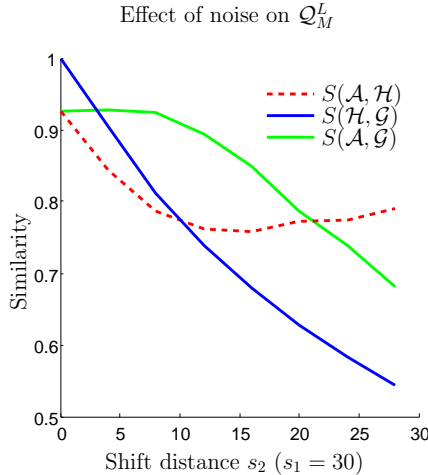
Figure 5. Effect of increasing noise on $\mathcal{Q}_M^L$. This plot shows the average similarity of the foreground classification $\mathcal{A}$ to ground truth $\mathcal{G}$ for the localized adaptive metrics $\mathcal{Q}_M^L$ on the rendered-object dataset using the multi-prototype setup (5 repetitions on 25 objects and 700 views of the dataset). The network was adapted with increasingly noisy hypotheses $\mathcal{H}$, which were obtained by scrambling with $s_1 = 30$ and increasing shift distance s_2. The capability of $\mathcal{Q}_M^L$ to approximate the ground truth information shows a graceful degradation with increasing noise. While keeping the prototypes nearly constant during the adaptation on a single image, the metrics adaptation is capable of obtaining a classifier for figure and ground which generalizes to the relevant object regions $S(\mathcal{A}, \mathcal{G}) > S(\mathcal{H}, \mathcal{G})$. Due to classification errors of the algorithm itself, the largest gain is achieved for intermediate levels of noise. A further increase of noise results in a learning of the hypothesis $S(\mathcal{A}, \mathcal{H})$, because the proportion of the object region is significantly reduced.

98% for $\mathcal{Q}_M^L$. The reason is a large share of correct background classification versus figure. Therefore the quality of the foreground classification is hard to assess from the measure D. Finally, because we use the ground truth data $\mathcal{H} = \mathcal{G}$ for supervised learning in this experiment the results can be considered as upper bounds of the foreground classification performance using the given setup.

3.4 Hypothesis refinement

On the basis of the preceding results we investigate the generalization capabilities of $\mathcal{Q}_M^L$ to the ground truth data. That is, the robustness against the increasing noise and the refinement of the initial hypothesis indicated by $S(\mathcal{A}, \mathcal{G}) > S(\mathcal{H}, \mathcal{G})$. Therefore we train a $\mathcal{Q}_M^L$ network by using multiple levels of distortions of $\mathcal{H}$ (Tab. 5). Because of classification errors introduced by the method itself (also observable in Tab. 1), some amount of distortion is required to observe the hypothesis-refinement effect for our scenario. In this

case, the higher model complexity enables a higher capability to generalize to the consistent parts of the object also in the presence of the increasing noise. Increasing the model complexity normally introduces the problem of over-fitting. Therefore we also compare the similarity $S(\mathcal{A}, \mathcal{H})$ of the foreground classification to the data used for training $\mathcal{H}$. We can observe that in particular for intermediate levels of noise, the foreground classification is more similar to the ground truth data than to the hypothesis, which indicates the good generalization capabilities. In the next section we will verify these observations on real image data recorded from the object learning scenario.

4 Object recognition scenario

We want to investigate the effect of the object segmentation derived by prototype and metrics adaptation on the data recorded in an online object recognition scenario. We use the data from [14] consisting of 50 natural, view centered objects with 300 training and 100 testing images without ground truth information. From the available depth and skin information the hypothesis $\mathcal{H}$ is computed without additional prior information on object position (as used in [13], see Sec. 2). In Comparison to the statistics of our dataset of rendered objects (Sec. 3.1.1), the average color of the foreground and background is (141, 119, 106) and (112, 99, 99) respectively. Similarly, the standard deviation in all feature channels is approximately 50 on both regions. Slightly larger, the foreground object occupies 22% of the image on average (39% for the bounding box).

To compare the results of the different methods where ground truth information is not available, the image regions defined by the foreground classification (i.e. the presented objects) are fed into a hierarchical feature processing stage [14]. For object learning and recognition, a separate nearest neighbor classifier is applied to the derived high dimensional shape features (Fig. 1). The resulting foreground segmentation is indirectly compared via the object classification performance of the nearest neighbor classifier on top of the segmented object views. Figure 6 shows samples for $\mathcal{A}$ and the recognition performance from using the depth-map itself, the hypothesis $\mathcal{H}$, the ASDF (used from [14]), and the results of the compared GLVQ-extensions. To distinguish between metrics and prototype learning, $\mathcal{Q}$(a) was trained with fast ($\alpha = 100$) and $\mathcal{Q}$(b) with slow learning rate ($\alpha = 0.05$). $\mathcal{Q}$ with adaptive metrics was trained analogously to Sec. 3 with $\alpha = 0.05, \beta = 0.005$ primarily adapting the metrics. While $\mathcal{Q}$ is not able to cope with the noisy supervised data, $\mathcal{Q}_M^L$ is capable of representing figure and ground on the basis of the most relevant features/feature combinations, which enables a correct foreground classification of the main object parts. Using foreground classifications of $\mathcal{Q}_M^L$ causes a significant improvement in recognition performance on real world data. Though

Object	Depth	$\mathcal{H}$	ASDF	$\mathcal{Q}(a)$

Perf.:	0.735	0.755	0.778	0.364

$\mathcal{Q}(b)$	$\mathcal{Q}_V^G$	$\mathcal{Q}_M^G$	$\mathcal{Q}_V^L$	$\mathcal{Q}_M^L$

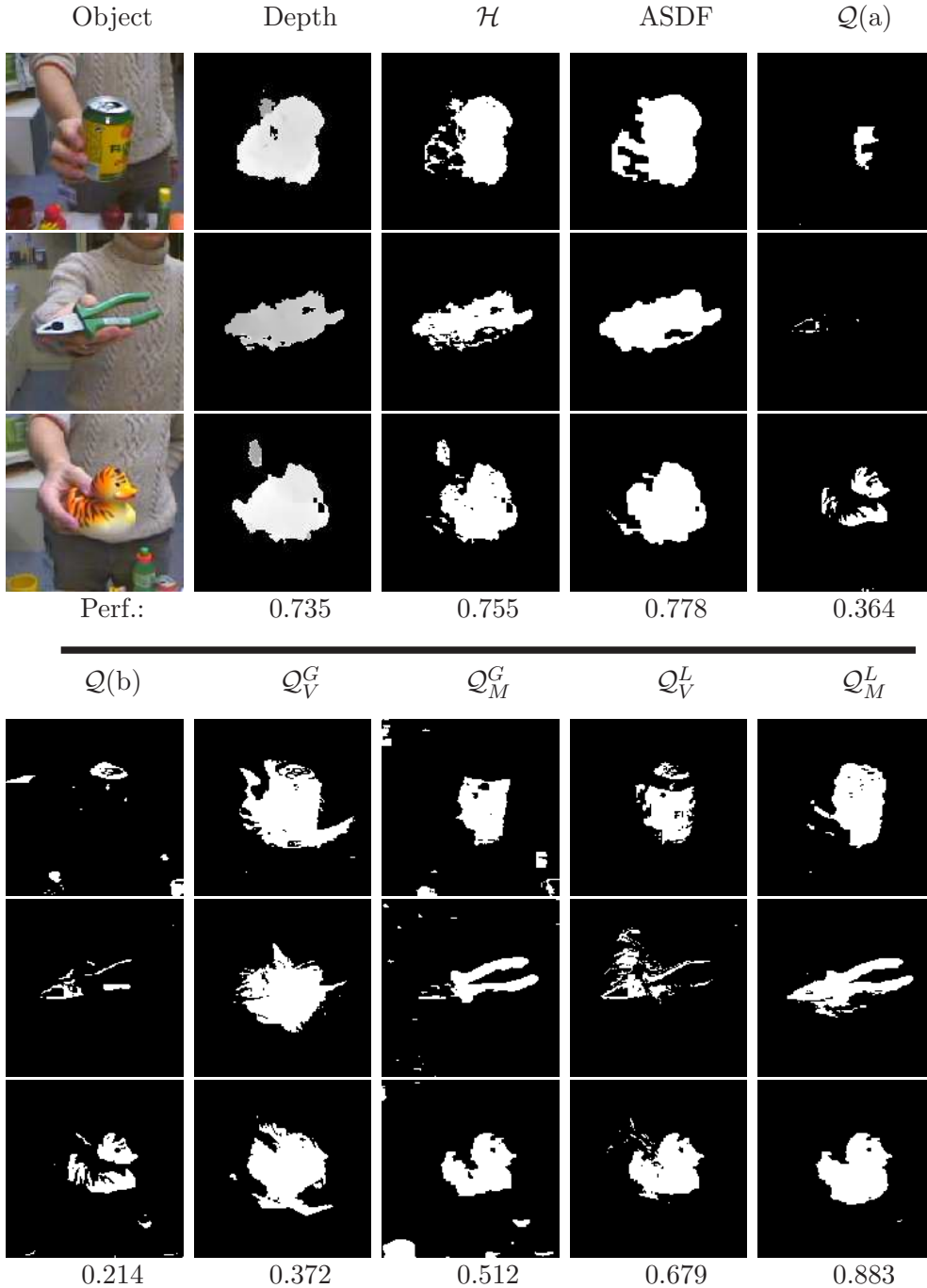| 0.214	0.372	0.512	0.679	0.883

Figure 6. From left to right: input image, depth-map, hypothesis $\mathcal{H}$ and derived $\mathcal{A}$ using $\mathcal{Q}$ with Euclidian and adaptive metrics. $\mathcal{Q}(a)$ uses a higher learning rate of $\alpha = 100$. Bottom row, the average object recognition performance of a separate nearest neighbor classifier on the high-dimensional shape features derived using the topographic visual hierarchy applied to the segmented object images (3 repetitions on 300 images for training, 100 for testing). We observe a gradual increase of segmentation quality and performance with increasing complexity of the metrics adaptation as well as the usage of local transformations rather than global ones.

the more complex metrics adaptations induces a higher computational load, the proposed segmentation method is still running at reasonable time for online learning of 7 frames/sec on this dataset (using a single core of a 2.66 GHz Intel Xeon processor machine).

5   Comparision to Graph-Cut segmentation

In the previous sections we showed that modeling figure and ground with prototypical feature representatives can strongly benefit from the localized metrics adaptation. Despite of a supervised learning method on the noisy hypotheses, the generalization capabilities can achieve a large gain in segmentation quality. While we are mainly concerned with an online application, this method is not constrained to that specific scenario, as long as the hypothesis is provided. Particularly interesting for an application of this method is the dependence of the generalization capability on the model complexity, the properties of the derived relevance factors as well as a comparison with the capability of other models. Prominent state of the art methods for segregating single objects from the backgrounds are methods based on Level-Sets [11] and Markov Random Fields [23,9]. To allow a comparison with these methods we apply the proposed method on the dataset [1] introduced by Rother et al. [9], which was also used for Level-Sets [11]. To our knowledge, currently the Graph-Cut [23] segmentation achieves the best performance on this dataset. For this benchmark the ground truth information is available, which allows a quantification of the segmentation quality. Furthermore, for each of the images a grey-value image called Trimap is available. This map mimics a user interaction that can provide hints to the algorithm about the relation of every pixel to figure or ground, encoded by Trimap $\in \{0, 64, 256\}$. Furthermore Trimap $\in \{128\}$ encodes for unknown status.

The benchmark dataset consists of quite different images. While there are many images with homogeneous object and/or background regions, other scenes are more difficult. In principle, prototype based methods are confronted with a model selection problem, that is, to determine the appropriate number of prototypes for each class. We decide to apply the $\mathcal{Q}_M^L$ method with the following setup. The number of prototypes is investigated with two different settings, consisting of one prototype for each class for the first setup, respectively two prototypes for the second setup. This might be insufficient for some of the images, but on the other hand increasing this number leads to overfitting effects on the simpler scenes and impairs the overall performance as well. Using only one prototype for each class further allows for a more

---

[1]   http://research.microsoft.com/vision/cambridge/i3l/segmentation/GrabCut.htm

Method	Error rate (avg. and std. dev.)
$\mathcal{H}$	07.72% ± 03.41
$\mathcal{Q}_M^L$, 2 prototypes, unconstrained Bimap	04.42% ± 03.04
$\mathcal{Q}_M^L$, 4 prototypes, unconstrained Bimap	04.15% ± 03.15
Graph-Cut, Trimap	02.38% ± 01.51
Graph-Cut, contrained Bimap	04.76% ± 04.36
Graph-Cut, unconstrained Bimap	12.90% ± 12.70

Table 2

Comparison of the error rates from $\mathcal{Q}_M^L$ and Graph-Cut applied to the benchmark dataset. Here the pixel-wise error rates $(1 - D(\mathcal{A}, \mathcal{G}) * 100)$ are used to achieve comparability to the cited literature [9,11]. First the hypothesis itself is evaluated, and then the metrics learning was applied with a two-prototype and four-prototype setup. For Graph-Cut several settings where used, which differ in the usage of the information provided from the Trimaps. While Graph-Cut strongly relies on this information to achieve a good performance, the proposed method is capable to cope with the unconstrained setting. Increasing the model complexity to multiple prototypes can increase the performance on the more complex samples of the dataset.

detailed inspection of the derived relevance factors over multiple repetitions. The learning rates are fixed with $\alpha = 0.05$ and $\beta = 0.005$. Due to the significantly higher image dimensions and the applications on single images, a larger number of 500000 trainings steps for each image is performed. The hypothesis to train the classifier for figure and ground is derived from the provided Trimaps of the database. That is, to train the LVQ network, all pixels whose corresponding values of the Trimap $\in \{128, 256\}$ are used as training data for foreground and otherwise for background.

In Table 2 the error rates of the derived foreground segmentations are compared to the results of a Graph-Cut [23] implementation. The parameters of the Graph-Cut model are λ, which is set to $1/15$ in all experiments, and σ. While λ specifies a relative importance of the region properties in the error functional of the Markov Random Field, σ is part of the boundary property term which defines cost for cutting the edge between two neighboring pixels. The parameter σ is estimated from the data as proposed in [9]. Like Graph-Cut, the proposed method is capable to derive a figure ground segregation that improves the initial guess. A significant difference between the proposed method and Graph-Cut is the large variance of the results. This variance occurs between the different images (Tab. 2), as well as for multiple repetitions on the same image, visible from the relevance factors in Fig. 8. This can be explained by the usage of the parameter from the online learning setup on this database. For gradient descent convergence to local a minimum is guaranteed, which depends on the initialization. Therefore the purely random initialization of the prototypes as well as the constant learning rate have a significant
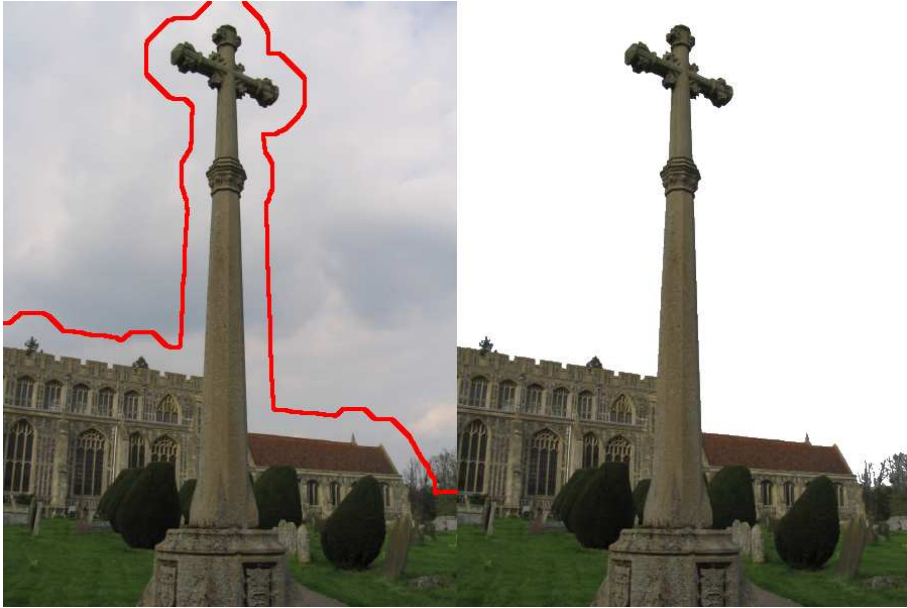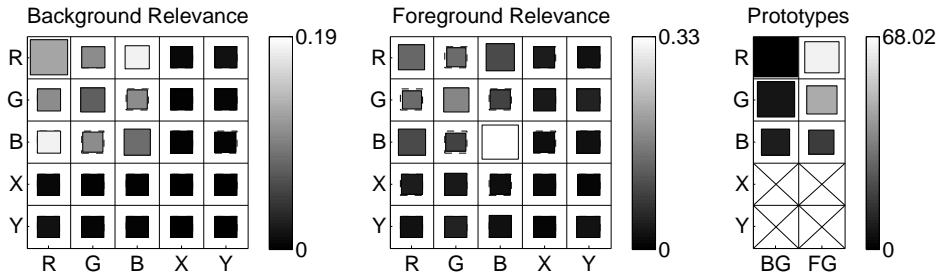
Example 1

Example 2

Figure 7. Sample segmentations derived by $\mathcal{Q}_M^L$ with the 2 prototype setup, one for foreground and one for background. The pixels used for the training of the foreground class are bounded by the red line on the input images (left). The right image is an overlay of the input image with the resulting foreground segmentation.

impact. A more sophisticated initialization as well as decreasing learning rate over time might relieve such effects. The important difference stems from the observation that the performance of the Graph-Cut approach in drastically decreased if the full information of the Trimap is not used. Obviously the Graph-Cut methods strongly rely on information which parts of the scene are definitely foreground (Trimap $\in \{256\}$) or background (Trimap $\in \{0, 64\}$) which is used as hard constraint for the algorithm. The proposed method does not rely on this information and uses the unconstrained Bimap. That is, the whole image is used to build the models and all pixels have to be classified afterwards and can be changed in their assignment to foreground or background. An intermediate setting where only the background is used as hard constraint is referred as "constrained Bimap".

Relevance Factors for Example 1
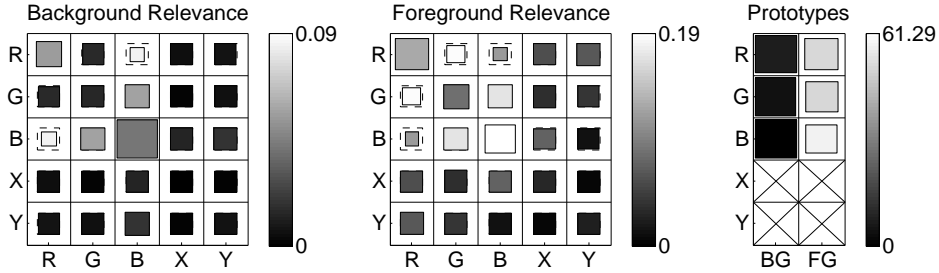

Relevance Factors for Example 2

Figure 8. Visualization of the derived relevance factors and their standard deviation on examples of Figure 7. The size of the boxes encodes the magnitude of the weights (empty box = -1, full box = 1). A middle sized box represents a value of 0, also visualized by the dotted inner square if this factor is smaller than zero. The intensity encodes for the standard deviation according to the right color encoding. Rightmost, the prototypes are visualized in a similar manner, where the size of the box encodes the value between zero and maximum. To ease the visualization, the position features are ignored because of their different range. The color of the boxes encodes for the standard deviation as before. These plots visualize the variance of the resulting relevance matrixes dependent on the initialization of the prototypes. Compared to standard prototype based representations, the properties of the represented image regions are primarily reflected in the relevance factors rather than in the prototypes.

The usage of only one prototype for each class allows an evaluation of the relevance learning over multiple repetitions on the same image. In Fig. 8 the average relevance matrix derived by $\mathcal{Q}_M^L$ and the standard deviation are visualized on the basis of 10 repetitions, visualized for example 1 and example 2 in Fig. 7. Firstly, the variance of the results is not equally distributed on all relevance factors, which means that they are the result of a systematic process converging into the local minima of the error function depending on the initialization. Secondly, the derived relevance factors reflect the image data in their prominent colors. For example the color blue gets a high weight for the foreground in example 1 and a large weight for background in example 2. Compared to standard prototype based learning, we observe that the properties of the image region represented by the pair of prototype and relevance matrix are primarily reflected in the relevance factors. This reflects the capabilities

Method	Error rate, noisy $\mathcal{H}$	Error rate, $\mathcal{H} = \mathcal{G}$
$\mathcal{H}$	$5.64\% \pm 1.00$	$0.00\% \pm 0.00$
Graph-Cut	$8.30\% \pm 8.82$	$7.30\% \pm 7.92$
$\mathcal{Q}_M^L$, 2 prototypes	$5.48\% \pm 4.18$	$4.87\% \pm 3.66$
$\mathcal{Q}_M^L$, 20 prototypes	$1.65\% \pm 1.03$	$1.32\% \pm 0.01$

Table 3
Comparison of metrics adaptation with Graph-Cut on the dataset described in Sec. 3.1.1. Here the average and standard deviation of the pixel-wise error rates similar to Tab. 2 are used. The metrics adaptation was applied with two different setups. The first setup consists of 20 prototypes and is described in Sec. 3.1.1, for the second setup only 2 prototypes are used to allow a comparison to Tab. 2. Metrics learning as well as Graph-Cut is applied on the ground truth data and a scrambled hypothesis, where the unconstrained Bimap setting was used in all conditions. Given an appropriate model complexity, the prototype-based learning is less sensitive to the quantity and quality of the provided training data.

of metrics adaptation to compensate for the reduced prototype adaptation, caused by the different learning rates. But in general the relevance cannot be rated on the region they present alone and in particular correlations between features are difficult to judge for human observers. Instead, the derived relevance factors are the results of the learning dynamics on foreground and background of the current image. Thus they cannot be simply transferred from one image to another.

5.1 Comparison to Graph-Cut on the dataset of rendered objects

Despite of the different usage of the information provided by Trimaps, another quite important difference is the usage of histograms to model figure and ground in the Graph-Cut approach. This relies on sufficient image data to model the feature distributions. On can expect that the performance depends on the image size where prototypical representatives yield a more compact model of the image data and can cope with smaller image dimensions. To evaluate the performance of the Graph-Cut method in our scenario we use our database of rendered images. Another problem that does not occur on the benchmark database is that the hypothesis is allowed to have holes. That is, in particular for the depth information estimated on homogeneous surfaces, the hypothesis for foreground can consists of regions where no measurement of depth is available (pixels that can be ignored to train the models) or simply assigned to background as it is the case for the rendered image database. Therefore we cannot use the Trimap or constrained Bimap setup for Graph-Cut, without further preprocessing (e.g. compute the convex hull of the hypothesis). Using the unconstrained Bimap setup as well as the histograms on

single small images strongly impairs the performance of the Markov Random Field approach. The prototype based approach shows a significantly stronger robustness under these conditions. Nevertheless, a two prototyp setup on this scenario does not have the appropriate model complexity. The setup optimized on this scenario allows a significant improvement of figure-ground segregation comparable to the results of Tab. 1 and Fig. 5.

6 Conclusion

In this paper, we propose a fast image segmentation scheme which is capable of refining a given hypothesis for arbitrary background conditions. We model figure and ground by prototypical feature representatives and compare several metrics extensions applied to GLVQ to improve this approach. Finally, we adopt LGMLVQ in the domain of figure-ground segregation for this purpose. In comparison to other metrics (Sec. 3.3, 4), we have shown that the extension to local matrices of relevance vectors leads to improved foreground classification resulting in a significant enhancement of object learning and recognition. Compared to the ASDF approach [13], which also directly addresses the foreground segmentation from an initial hypothesis, the supervised learning does not rely on additional a priori assumptions about object position, size and segment-selection. In comparison with a current state of the art object segmentation method, we show that the proposed method has fewer constraints on the provided training data and is less sensitive to the quality of the initial hypothesis.

To explain the positive effect on hypothesis refinement, the number of prototypes and the introduction of the pixel position as additional features are important. The number of prototypes is constrained to be small and therefore the algorithm is forced to represent the most dominant structures in the image by means of this limited set. Important for interpreting the capabilities on hypothesis refinement is the fact that the noise induced by a wrong hypothesis is not randomly distributed over the image, but structured near the corresponding object. This noise, as well as similar colors in foreground and background, is responsible for overlapping clusters in feature-space. Transferring this feature into a higher dimensional space by adding the position alone does not solve this problem. Only the non-linear decision boundaries introduced by local transformations in connection with the even higher flexibility by using multiple prototypes for each class allow a better representation of this heterogeneously structured data.

By optimizing the parameters to the most complex metrics adaptation we found that the largest benefit of metrics adaptation can be obtained by focusing the learning on this part. Adapting the prototypes very slowly allows us

24

to separate the effects of prototype and metrics adaptation and to compare the impact of several adaptive metrics by exclusively varying their complexity and none of the remaining parameters (learning rate, number of prototypes). Further we optimize the learning rate for GLVQ which allows us to compare the effects of prototype adaptation vs. metrics adaptation. We observe that increasing the complexity of the metrics successively increases the generalization capabilities and compensates for the missing prototype adaptation. Also metrics adaptation yields a clear advantage in particular on noisy supervised information.

The capability of optimizing the foreground classifier on a stable set of prototypes offers some interesting possibilities. In the experiments, the prototypes have been randomly initialized. In particular, the learning of the network is impaired by a fast prototype adaptation on partially inconsistent training data. Adapting only the metrics while keeping the prototypes stable yields the desired generalization capabilities. This motivates for future work to introduce a higher flexibility of the prototype adaptation by a separate learning method while using metrics adaptation to refine the hypothesis. To achieve this and to address the general model selection problem, the unsupervised Instantaneous Topological Map [21] offers the advantage to initialize the prototypes and estimate their number for each class [22]. As the proposed method is not constrained to a particular set of feature maps (e.g. RGB or other color spaces like CIE Lab or HSV), further investigation will also address the introduction of additional features (e.g. texture). The extension to a three class setup for a direct integration of the skin color detection seems promising, too.

References

[1] B. Leibe, A. Leonardis, B. Schiele, Robust object detection with interleaved categorization and segmentation, International Journal of Computer Vision 77 (2007) 259–289.

[2] S. X. Yu, J. Shi, Object-specific figure-ground segregation, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, IEEE Computer Society, 2003, pp. 39–45.

[3] E. Borenstein, E. Sharon, S. Ullman, Combining top-down and bottom-up segmentation, Conference on Computer Vision and Pattern Recognition Workshop (CVPRW) Vol. 4 (2004) 46.

[4] E. Borenstein, S. Ullman, Learning to segment, in: European Conference on Computer Vision (ECCV), LNCS, Springer, 2004, pp. 315–328.

[5] M. P. Kumar, P. H. S. Torr, A. Zisserman, OBJ CUT, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, IEEE Computer Society, 2005, pp. 18–25.

[6] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 888–905.

[7] H. Wersing, J. J. Steil, H. Ritter, A competitive layer model for feature binding and sensory segmentation, Neural Computation 13 (2) (2001) 357–387.

[8] S. Weng, H. Wersing, J. J. Steil, H. Ritter, Learning lateral interactions for feature binding and sensory segmentation from prototypic basis interactions, IEEE Transactions Neural Networks 17 (4) (2006) 843–862.

[9] C. Rother, V. Kolmogorov, A. Blake, "GrabCut": interactive foreground extraction using iterated graph cuts, ACM Transactions on Graphics 23 (3) (2004) 309–314.

[10] J. Sun, W. Zhang, X. Tang, H. Shum, Background Cut, in: European Conference on Computer Vision, Springer, 2006, pp. II: 628–641.

[11] D. Weiler, J. Eggert, Multi-dimensional histogram-based image segmentation, in: Proceedings of the 14th International Conference on Neural Information Processing (ICONIP), Springer, 2007.

[12] R. Achanta, F. Estrada, P. Wils, S. Süsstrunk, Salient region detection and segmentation, in: A. Gasteratos, M. Vincze, J. K. Tsotsos (Eds.), Computer Vision Systems, Vol. 5008 of LNCS, Springer, 2008, pp. 66–75.

[13] J. J. Steil, M. Götting, H. Wersing, E. Körner, H. Ritter, Adaptive scene-dependent filters for segmentation and online learning of visual objects, Neurocomputing 70 (7-9) (2007) 1235–1246.

[14] H. Wersing, S. Kirstein, M. Götting, H. Brandl, M. Dunn, I. Mikhailova, C. Goerick, J. J. Steil, H. Ritter, E. Körner, Online learning of objects in a biologically motivated visual architecture, International Journal of Neural Systems 17 (4) (2007) 219–230.

[15] A. Sato, K. Yamada, Generalized learning vector quantization, in: Advances in Neural Information Processing Systems, Vol. 7, 1995, pp. 423–429.

[16] B. Hammer, T. Villmann, Generalized relevance learning vector quantization, Neural Networks 15 (8-9) (2002) 1059–1068.

[17] P. Schneider, M. Biehl, B. Hammer, Relevance matrices in LVQ, in: M. Verleysen (Ed.), Proceedings of the European Symposium on Artificial Neural Networks (ESANN), d-side publications, 2007, pp. 37–42.

[18] J. Fritsch, S. Lang, M. Kleinehagenbrock, G. A. Fink, G. Sagerer, Improving adaptive skin color segmentation by incorporating results from face detection, in: 11th IEEE International Workshop on Robot and Human Interactive Communication (ROMAN), IEEE, 2002, pp. 337–343.

[19] P. Schneider, M. Biehl, F.-M. Schleif, B. Hammer, Advanced metric adaptation in Generalized LVQ for classification of mass spectrometry data, in: Proceedings of 6th International Workshop on Self-Organizing Maps (WSOM), 2007, published on CD (Univ. Bielefeld 2007).

[20] M. Biehl, B. Hammer, P. Schneider, Matrix learning in learning vector quantization, Technical Report, Insitute of Informatics, Clausthal University of Technology (2006).

[21] J. Jockusch, H. Ritter, An instantaneous topological mapping model for correlated stimuli, in: Proceedings of the International Joint Conference on Neural Networks (IJCNN 99), 1999, p. 445.

[22] A. Denecke, Anwendung vektorbasierter Netzwerke zur adaptiven Segmentierung von Bildfolgen, Master's thesis, University of Bielefeld, Faculty of Technology (2005).

[23] Y. Y. Boykov, M. P. Jolly, Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images, in: Eighth International Conference on Computer Vision (ICCV'01), Vol. 1, 2001, pp. 105–112.