Pattern Anal Applic (2010) 13:333–348
DOI 10.1007/s10044-009-0163-0

THEORETICAL ADVANCES

A system for 3D texture-based probabilistic object recognition and its applications

Marcin Grzegorzek

Received: 7 March 2008 / Accepted: 8 April 2009 / Published online: 1 July 2009

Abstract This article presents a system for texture-based probabilistic classification and localisation of three-dimensional objects in two-dimensional digital images and discusses selected applications. In contrast to shape-based approaches, our texture-based method does not rely on object features extracted using image segmentation techniques. Rather, the objects are described by local feature vectors computed directly from image pixel values using the wavelet transform. Both gray level and colour images can be processed. In the training phase, object features are statistically modelled as normal density functions. In the recognition phase, the system classifies and localises objects in scenes with real heterogeneous backgrounds. Feature vectors are calculated and a maximisation algorithm compares the learned density functions with the extracted feature vectors and yields the classes and poses of objects found in the scene. Experiments carried out on a real dataset of over 40,000 images demonstrate the robustness of the system in terms of classification and localisation accuracy. Finally, two important real application scenarios are discussed, namely recognising museum exhibits from visitors' own photographs and classification of metallography images.

Keywords Object recognition · Statistical modelling · Wavelet analysis · Image processing

M. Grzegorzek (✉)
Information Systems and Semantic Web Research Group,
University of Koblenz-Landau, Universitaetstr. 1,
56070 Koblenz, Germany
e-mail: marcin@uni-koblenz.de

1 Originality and contribution

The system described in the present paper is able to successfully deal with tasks from real world environments including real heterogeneous background of high complexity, varying illumination, and object occlusions. Therefore, it has been used in two important real application scenarios, namely recognising museum exhibits from visitors' own photographs and classification of metallography images. In the first scenario, we envisage the museum offering a Web 2.0 service that automatically annotates user's photos and offers added value content that can be personalised to the users based on their interest in specific artefacts. At some point after a visit, the user submits a set of digital photos taken inside the museum to the museum's site. The artefacts depicted are recognised using the approach presented in this paper providing a link between the users content and the museum's own data archives. The service then tags the user's photo with some metadata related to the artefact in question (e.g. name, date and location of origin, etc.). In the second scenario we analyse metallography image from the Ironworks in Ostrava (Czech Republic). The aim of this analysis is monitoring the quality process in the steel plant.

2 Introduction

A fundamental problem of *computer vision* is the recognition of objects in digital images. The term *object recognition* covers both *classification* and *localisation*. Classification and localisation of objects in images is a useful, and often indispensable step, for many real-life computer vision applications. Algorithms for automatic computational object recognition can be applied in areas

such as face classification [9, 38], fingerprint classification [42, 30], handwriting recognition [5, 13], service robotics [41, 43], medicine [3, 20], visual inspection [17, 29], the automobile industry [6, 7], etc. Although successful applications have been developed for some tasks, e.g., fingerprint classification, there are still many other areas that could potentially benefit from object recognition. The system described in this article has been tested in real application scenarios. One of these is the recognition of museum exhibits from visitors' personal photographs, another is the analysis of metallography images from an ironworks.

For the problem of object classification, the system is supposed to determine the classes of objects occurring in an image from the set of known object classes $\Omega = \{\Omega_1, \Omega_2, \ldots, \Omega_\kappa, \ldots, \Omega_{n_\Omega}\}$. However, the number of objects in a scene is typically unknown and must also be determined. In the case of object localisation, the recognition system must estimate the pose of an object in the image. The object pose is defined with a translation vector $\boldsymbol{t} = (t_x,\ t_y,\ t_z)^{\mathrm{T}}$ and three rotation angles (ϕ_x, ϕ_y and ϕ_z) around the axes of the Cartesian coordinate system. The origin of the Cartesian coordinate system is placed in the symmetrical centre of the image, the x- and y-axes lie in the image plane, and the z-axis is orthographic to the image plane (see Fig. 1). When the object size and appearance do not change in a given image, the object pose is defined with internal transformation parameters $\boldsymbol{t} = (t_x,\ t_y,\ t_z)^{\mathrm{T}}$ and $\phi_{\mathrm{int}} = \phi_z$ only. This two-dimensional (2D) case can be seen in the first row of Fig. 1. A more realistic and challenging situation occurs when the object is transformed with external pose parameters $\boldsymbol{t}_{\mathrm{ext}} = t_z$ and $\boldsymbol{\phi}_{\mathrm{ext}} = (\phi_x, \phi_y)^{\mathrm{T}}$. In the second row of Fig. 1 the object changes not only in size but also in appearance for some points of view. The goal of object localisation is to determine both the internal and the external object pose parameters in a real environment.

For the recognition of three-dimensional (3D) objects in 2D images, two main approaches are known in computer vision: based on the result of object segmentation (shape-based), or by directly using the object texture (texture-based). Shape-based methods make use of geometric features such as lines or corners extracted by segmentation operations. These features and their relationships are then used for object description [4, 14, 16, 18]. However, the segmentation-based approach often suffers from errors due to loss of image details or other inaccuracies resulting from the segmentation process. Texture-based approaches avoid these disadvantages by directly using the image data, i.e., the pixel values, without a previous segmentation step [26, 32, 34]. For this reason the texture-based method for object recognition has been chosen to develop the system presented in this contribution.

The rest of the article is structured as follows. Section 3 reviews the state of the art in the area of object recognition. Section 4 describes the training procedure for object

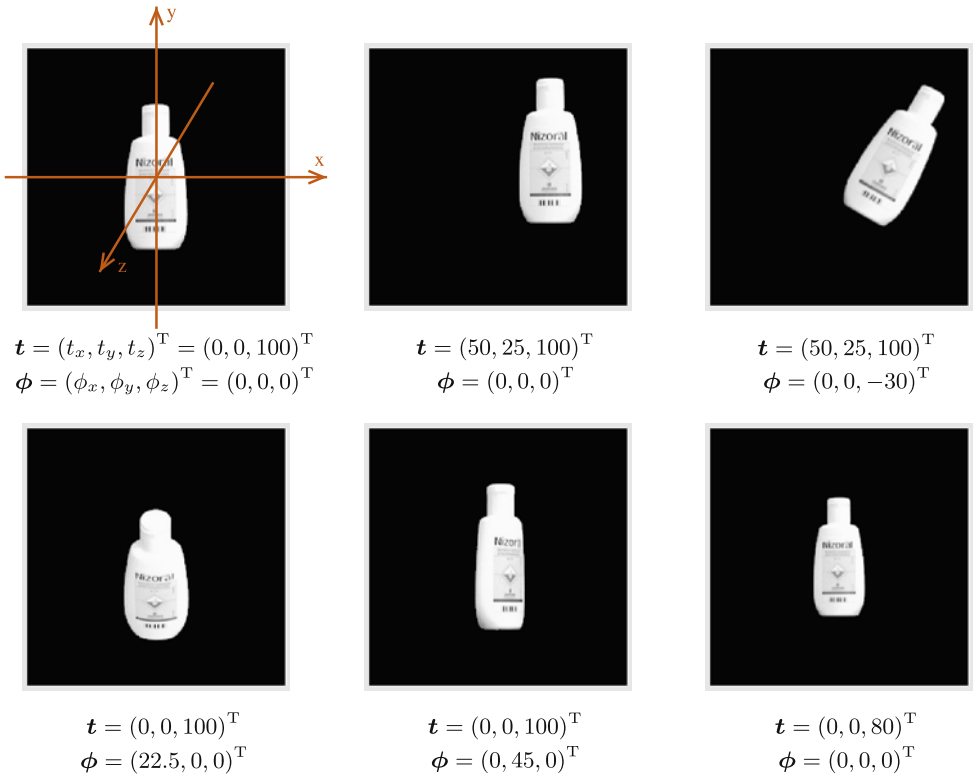

Fig. 1 Examples of object poses and their values. The components of the internal translation vector $\boldsymbol{t}_{\mathrm{int}} = (t_x, t_y)^{\mathrm{T}}$ are given in pixels, the components of the rotation vector $\boldsymbol{\phi} = (\phi_x, \phi_y, \phi_z)^{\mathrm{T}}$ in degrees (°), and the external translation (scaling) $t_{\mathrm{ext}} = t_z$ in percent (%) of a reference object size (*top left*)

localisation and classification. The object recognition phase is detailed in Sect. 4. Section 5 covers the experimental results achieved on a large database of over 40,000 images of real objects captured against heterogeneous backgrounds. Section 6 describes two real application scenarios successfully implemented with our system: classification of museum artefacts and classification of metallography images. Conclusions and directions for future work are presented in Sect. 7.

3 Related approaches

In this section, related approaches selected from the wide area of object recognition are presented. Moreover, those features of our method which ensure its novelty and originality are pointed out at the end of the section.

Amit et al. [1] proposes an algorithm for multi-class shape detection in the sense of recognising and localising instances from multiple shape classes. This approach is formulated as a two-step process in which local indexing primes global interpretations. During indexing a list of instantiations (shape identities and poses) is compiled constrained only by no missed detections at the expense of false positives. Global information, such as expected relationships among poses, is incorporated afterward to remove ambiguities.

In Lowe [23], a method for extracting distinctive invariant features from images that can be used to perform reliable matching between different views of an object or scene is presented. The features are invariant to image scale and rotation, and are shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination. The recognition proceeds by matching individual features to a database of features from known objects using a fast nearest-neighbour algorithm, followed by a Hough transform to identify clusters belonging to a single object, and finally performing verification through least-squares solution for consistent pose parameters.

Support vector machines (SVMs) have been recently proposed as a new technique for pattern recognition. Intuitively, given a set of points which belong to either of two classes, a linear SVM finds the hyperplane leaving the largest possible fraction of points of the same class on the same side, while maximising the distance of either class from the hyperplane. The hyperplane is determined by a subset of the points of the two classes, named support vectors, and has a number of interesting theoretical properties. In Pontil and Verri [31], linear SVMs for 3D object recognition are used. The proposed system does not require feature extraction and performs recognition on images regarded as points of a space of high dimension without estimating pose.

The appearance of an object is composed of local structure. This local structure can be described and characterised by a vector of local features measured by local operators such as Gaussian derivatives or Gabor filters. In Schiele and Crowley [36] a technique is presented where appearances of objects are represented by the joint statistics of such local neighbourhood operators. As such, this represents a new class of appearance-based techniques for computer vision. Based on joint statistics, the paper develops techniques for the identification of multiple objects at arbitrary positions and orientations in a cluttered scene.

In Schneiderman and Kanade [37], a trainable object detector and its instantiations for detecting faces and cars at any size, location, and pose is described. To cope with variation in object orientation, the detector uses multiple classifiers, each spanning a different range of orientation. Each of these classifiers determines whether the object is present at a specified size within a fixed-size image window. To find the object at any location and size, these classifiers scan the image exhaustively. Each classifier is based on the statistics of localised parts, whereas each part is a transform from a subset of wavelet coefficients to a discrete set of values.

In Torralba et al. [39], the problem of detecting a large number of different classes of objects in cluttered scenes is taken into consideration. A multi-task learning procedure is proposed, based on boosted decision stumps, that reduces the computational and sample complexity, by finding common features that can be shared across the classes (and/or views). For a given performance level, the total number of features required, and therefore the run-time cost of the classifier, is observed to scale approximately logarithmically with the number of classes.

Jin and Geman [15] proposes a mathematical framework (a "composition machine") for constructing probabilistic hierarchical image models, designed to accommodate arbitrary contextual relationships. Moreover, it describes a demonstration system for reading Massachusetts license plates in an image set collected at Logan Airport. This demonstration system detects and correctly reads more than 98% of the plates, with a negligible rate of false detection.

In order to compare different methods for object recognition, in Leibe and Schiele [19] a new database specifically tailored to the task of object categorisation is presented. It contains high-resolution colour images of 80 objects from 8 different categories, for a total of 3,280 images. It is used to analyse the performance of several appearance- and contour-based methods. The best categorisation result is obtained by an appropriate combination of different methods.

In Lowe [22], an object recognition system is described that uses a new class of local image features. The features are invariant to image scaling, translation, and rotation, and partially invariant to illumination changes and affine or 3D projection. These features share similar properties with neurons in inferior temporal cortex that are used for object recognition in primate vision. Image keys are created that allow for local geometric deformations by representing blurred image gradients in multiple orientation planes and at multiple scales. The keys are used as input to a nearest-neighbour indexing method that identifies candidate object matches.

In Mahamud and Hebert [24], a multi-class object detection framework whose core component is a nearest neighbour search over object part classes is presented. The optimal distance measure is modelled using a linear logistic model that combines the discriminative powers of more elementary distance measures associated with a collection of simple to construct feature spaces such as colour, texture and local shape properties.

As can be seen, a lot of valuable research work has been done in the area of object recognition in the last years. However, some features of our system prove its novelty and originality as well as high performance in the sense of classification and localisation accuracy. Following aspects count to the strongest of our system:

- *Fusion of multiple views*

A new approach for the fusion of multiple views based on a recursive density propagation method is introduced. In contrast to passive algorithms, where the decision about class and pose of an object has to be taken based on one image, here more images are used. The additional images are used to gain more information about the scene and the observed objects. The experimental results show that the fusion improves the recognition rates substantially, especially for difficult conditions such as heterogeneous background within real world environments.

- *Fast training*

A new approach is proposed, where the image acquisition for the training phase is done with a hand-held camera. The poses of the objects in all training frames are computed using a structure-from-motion algorithm [12]. The whole learning process is, therefore, independent on environment assumptions, but we have to deal with an additional training inaccuracy.

- *Resolution level combination of wavelet transformation*

Feature extraction on three different resolution levels of the wavelet transformation is introduced, and three statistical object models for each object class are created in the training phase. The algorithm for object classification and localisation uses a combination of the object models obtained for these different resolution levels, which significantly improves the recognition rates.

- *Colour modelling*

In many research works, object models are created based on gray-level images. In the present work it is proposed to use the colour information of objects.

- *Multi-object scenes with context modelling*

In many research works, multi-object scenes are considered without attention to the context dependencies. In the present work, context modelling for multi-object scenes is introduced.

- *Model interpolation between the training viewpoints*

For training, objects can be acquired only from a finite number of viewpoints. However, in the recognition phase it might be useful to localise them also between the training points of view. For this reason, an interpolation step in the training phase is performed. This ensures that the statistical object models obtained in training can be considered as continuous functions defined on the pose parameter domain.

4 Training

4.1 Training data collection

In order to capture training data, objects are put on a turntable that rotates to set angles, and training images are taken for each of these angles. The camera is fixed on a mobile arm that can move around the object. The turntable position produces information about the rotation ϕ_y of the object around the vertical y axis. The position of the camera relative to the object yields the object's rotation ϕ_x around the horizontal x axis. The object's scale (translation t_z along the z) can be set with the zoom parameter of the camera, or by moving the camera closer or further from the object. By modifying the camera parameters and position, images can be captured from all top and sidewise views of the object with known external pose parameters (ϕ_{ext}, t_{ext}) for each training image. The translation of the object in the image plane (internal translation) $\boldsymbol{t}_{\text{int}} = (t_x, t_y)^{\text{T}}$ as well the internal rotation $\phi_{\text{int}} = \phi_z$ can be determined after the acquisition process from the relative position of the object in the image. The object pose parameters are usually given relative to each other, as can be seen in Fig. 1. For each object class, one image is chosen as the reference image. The pose of an object in an image is understood as being the 3D transformation (rotation and translation) that maps that object into the reference image.

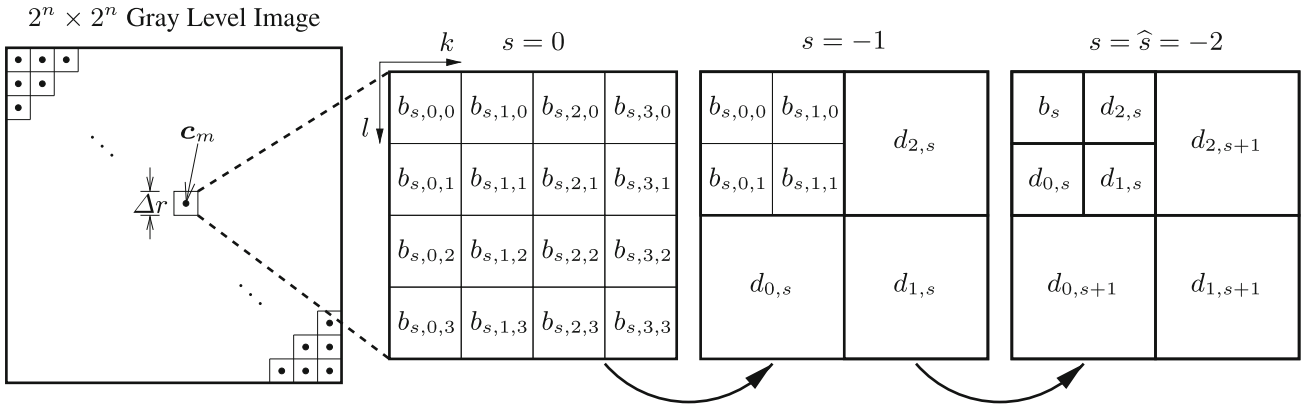

Fig. 2 2D signal decomposition with the wavelet transform for a local neighbourhood of size 4 × 4 pixels. The final coefficients result from gray values $b_{0,k,l}$ and have the following meaning: b_{-2} : low-pass horizontal and low-pass vertical, $d_{0,-2}$: low-pass horizontal and high-pass vertical, $d_{1,-2}$: high-pass horizontal and high-pass vertical, $d_{2,-2}$: high-pass horizontal and low-pass vertical

4.2 Feature extraction with the wavelet transform

Both gray-level and colour images can be used for object modelling. First, the system converts and resizes the original training scenes into gray-level or RGB images of size $2^n \times 2^n (n \in \mathbb{N})$ pixels, then local feature vectors in these images are computed. The main advantage of local feature vectors is that a local disturbance affects only the features in a surrounding area. In contrast, a global feature vector can completely change if only one pixel in the image varies. Moreover, it is easier to distinguish the object area from the background using local feature vectors.

The system determines a set of local feature vectors c_m for all preprocessed training images of an object via the discrete wavelet transform [25]. In order to calculate the c_m vectors, a grid with size $\Delta r = 2^{\widehat{|s|}}$, where $\widehat{s}$ is the minimum multiresolution scale parameter[1] s, is overlaid on the image [11]. Figure 2 depicts this procedure for the case of gray-level scenes divided into local neighbourhoods of size 4 × 4 pixels. Using the coefficients introduced in Fig. 2, the local feature vector c_m for the gray-level image is defined by,

$$c_m = \begin{pmatrix} c_{m,1} \\ c_{m,2} \end{pmatrix} = \begin{pmatrix} \ln\left(2^{\widehat{s}} |b_{\widehat{s}}|\right) \\ \ln\left[2^{\widehat{s}} \left(|d_{0,\widehat{s}}| + |d_{1,\widehat{s}}| + |d_{2,\widehat{s}}|\right)\right] \end{pmatrix}. \quad (1)$$

In the feature vector, the first component $c_{m,1}$ stores information about the mean gray-level (low-frequencies) in the local neighbourhood, while the second component $c_{m,2}$ represents discontinuities (high frequencies). The natural logarithm (ln) decreases the sensibility of the system to illumination changes and muffles any noises, which occur very often, especially in the real world environment. Its use is experimentally motivated in [34]. In the case of RGB images, each colour channel is treated independently. The feature computation for each channel is performed in the same way as for gray-level images (see Fig. 2). Therefore, the local feature vector for colour images has six components,

$$c_m = (c_{m,1}, c_{m,2}, c_{m,3}, c_{m,4}, c_{m,5}, c_{m,6},)^{\mathrm{T}}. \quad (2)$$

The first $c_{m,1}$ and the second $c_{m,2}$ components are calculated from the red channel, the third $c_{m,3}$ and the fourth $c_{m,4}$ from the green channel, and the fifth $c_{m,5}$ and the sixth $c_{m,6}$ from the blue channel [10]. Generally, the system is able to compute local feature vectors for any resolution scale $\widehat{s}$, but in practice $\widehat{s} \in \{-1, -2, -3\}$ is preferred.

4.3 Object area definition

Clearly, some feature vectors in each training image describe the object, while others belong to the background. In real life applications it cannot be assumed that the background is a priori known in the recognition phase. Therefore, only the feature vectors describing the object are considered for statistical object modelling. Since the object usually only composes a part of the image, a tightly enclosing bounding region O is defined for each object class. From here on we will term this bounding region the *object area*. Subsequently, the local feature vectors inside the object area are assigned to the object and termed *object feature vectors*, while the features outside this area O are denoted as *background feature vectors*. For the sake of clarity, we will use the term *object area* to actually refer to the set of features belonging to the object. The object area can change its location, orientation, and size from image to image depending on the object pose parameters. In the simplest case, when the object is rotated by $\phi_{\mathrm{int}} \in \mathbb{R}$ around the perpendicular axis to the image plane and translated by $\boldsymbol{t}_{\mathrm{int}} \in \mathbb{R}^2$ in the image plane, its appearance

[1] i.e., Further decomposition of the signal with the wavelet transform is not possible.

and size will not change. For more complex transformations in the external pose, not only its size, but also its appearance, i.e., pixel values in the object area, can change. Thus, for some external transformations $(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}})$ a local feature vector c_m describes the object $(c_m \in O)$, while for others the same vector belongs to the background $(c_m \notin O)$. For this reason, the object area is modelled as a function of the external pose parameters

$$O = O(\phi_{\text{ext}}, t_{\text{ext}}), \tag{3}$$

ideally within a continuous domain. This is done using the so-called *assignment functions* ξ defined for all feature vectors $\boldsymbol{c}_m$ and all training viewpoints $(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}})$ as,

$$\xi = \xi(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}). \tag{4}$$

The assignment function ξ_m decides, whether the feature vector $\boldsymbol{c}_m$ belongs to the object in the pose $(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}})$ or to the background, as follows,

$$\left\{ \begin{array}{l} \xi_m(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}) \geq S_O \Rightarrow \boldsymbol{c}_m \in O(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}) \\ \xi_m(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}) < S_O \Rightarrow \boldsymbol{c}_m \notin O(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}) \end{array} \right\}, \tag{5}$$

where the threshold value S_O is set experimentally and has the same value for all object classes. The assignment functions are trained for each training view separately

$$\xi_m(\boldsymbol{\phi}_{\text{ext}},\ t_{\text{ext}}) \left\{ \begin{array}{ll} 1, & \text{if } c_{m,1} \geq S_\xi \\ 2, & \text{if } c_{m,1} < S_\xi \end{array} \right\} \tag{6}$$

where S_ξ is a threshold value. Since $c_{m,1}$ results from a low-level filtering of a small neighbourhood it represents, the practical execution of the training is quite simple. The objects are taken on a nearly black background and a threshold value S_ξ decides whether they belong to the object or to the background. Since there is a finite number of training views $(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}})$, these are discrete functions initially, but after interpolation with the sine–cosine transformation they become continuous. Therefore, considering both the internal and external transformation parameters, the object area can be expressed by the function

$$O = O(\boldsymbol{\phi}, \boldsymbol{t}) \tag{7}$$

defined in a continuous six-dimensional pose parameter space $(\boldsymbol{\phi}, \boldsymbol{t})$.

4.4 Statistical object modelling

In order to handle illumination changes and low-frequency noise, the elements $\boldsymbol{c}_{m,q}$ of the local feature vectors c_m are interpreted as random variables. Assuming the object's feature vectors $\boldsymbol{c}_m \in O$ as statistically independent of the feature vectors outside the object area, the background feature vectors $\boldsymbol{c}_m \notin O$ can be disregarded and modelled separately as outlined in Sect. 3.5. The elements of the object feature vectors are represented with Gaussian density functions $p(\boldsymbol{c}_{m,q}|\mu_{m,q}, \sigma_{m,q}, \boldsymbol{\phi}, \boldsymbol{t})$. The mean $\mu_{m,q}$ and standard deviation $\sigma_{m,q}$ values are estimated for all training views $(\phi_{\text{ext}}, t_{\text{ext}})$, which form a subspace of $(\boldsymbol{\phi}, \boldsymbol{t})$. Assuming the statistical independence of the elements $\boldsymbol{c}_{m,q}$, which is valid due to their different interpretations in terms of signal processing (Sect. 3.2), the density function for the object feature vector $\boldsymbol{c}_m \in O$ can be written as,

$$p(\boldsymbol{c}_m|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m, \boldsymbol{\phi}, \boldsymbol{t}) = \prod_{q=1}^{N_q} p(c_{m,q}|\mu_{m,q}, \sigma_{m,q}, \boldsymbol{\phi}, \boldsymbol{t}), \tag{8}$$

where $\boldsymbol{\mu}_m$ is the mean value vector, $\boldsymbol{\sigma}_m$ the standard deviation vector, and N_q the dimension of the feature vector $\boldsymbol{c}_m$ ($N_q = 2$ for gray-level images, $N_q = 6$ for colour images). Further, it is assumed that the feature vectors belonging to the object $\boldsymbol{c}_m \in O$ are statistically independent. Under this assumption, an object can be described by the probability density p as follows,

$$p(O|\boldsymbol{B}, \boldsymbol{\phi}, \boldsymbol{t}) = \prod_{\boldsymbol{c}_m \in O} p(\boldsymbol{c}_m|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m, \boldsymbol{\phi}, \boldsymbol{t}). \tag{9}$$

where $\boldsymbol{B}$ comprises the mean value vectors $\boldsymbol{\mu}_m$ and the standard deviation vectors $\boldsymbol{\sigma}_m$. This probability density is termed the *object density* and, taking into account (8), can be written in more detail as,

$$p(O|\boldsymbol{B}, \boldsymbol{\phi}, \boldsymbol{t}) = \prod_{\boldsymbol{c}_m \in O} \prod_{q=1}^{N_q} p(c_{m,q}|\mu_{m,q}, \sigma_{m,q}, \boldsymbol{\phi}, \boldsymbol{t}), \tag{10}$$

In reality, neighbouring feature vectors might be statistically dependent, but considering the full neighbourhood relationship, e.g., as a Markov Random Field [35], leads to a very complex model. In order to complete the object description with the object density (10), the means $\mu_{m,q}$ and the standard deviations $\sigma_{m,q}$ for all object feature vectors $\boldsymbol{c}_m$ have to be learned. For this purpose, N_ρ training images of each object $\boldsymbol{f}_\rho$ are used in association with their corresponding transformation parameters $(\boldsymbol{\phi}_\rho, \boldsymbol{t}_\rho)$. The mean vectors $\boldsymbol{\mu}_m$ concatenated, written as $\boldsymbol{\mu}$, and the standard deviation vectors $\boldsymbol{\sigma}_m$ concatenated, written as $\boldsymbol{\sigma}$, can be estimated from the maximisation of the object density (10) over all N_ρ training images,

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\sigma}}) = \underset{(\boldsymbol{\mu}, \boldsymbol{\sigma})}{\operatorname{argmax}} \prod_{\rho=1}^{N_q} p(O|\boldsymbol{B}, \boldsymbol{\phi}_\rho, \boldsymbol{t}_\rho). \tag{11}$$

As a result of a subsequent interpolation step, the mean vectors $\boldsymbol{\mu}_m$ and standard deviation vectors $\boldsymbol{\sigma}_m$ are trained for all pose parameters $(\boldsymbol{\phi}, \boldsymbol{t})$ in a continuous sense.

4.5 Statistical background modelling

As mentioned in Sect. 3.4, the background feature vectors $\boldsymbol{c}_m \notin O$ are assumed to be statistically independent of the

feature vectors inside the object area O and can be modelled separately. Since in the recognition phase the background is a priori unknown, each possible value of the background feature vector element $\boldsymbol{c}_{m,q}$ can be observed with the same probability. Thus, they are modelled as uniform random variables, and their constant density functions,

$$p(c_{m,q}) = \frac{1}{\max(c_{m,q}) - \min(c_{m,q})} \tag{12}$$

do not depend on the transformation parameters $(\boldsymbol{\phi}, \boldsymbol{t})$. Assuming the statistical independence of $c_{m,q}$, (12) can be extended to

$$p(\boldsymbol{c}_m) = \prod_{q=1}^{N_q} \frac{1}{\max(c_{m,q}) - \min(c_{m,q})} = p_b, \tag{13}$$

where p_b is a constant value called *background density*.

4.6 Statistical context modelling

Usually, statistical approaches for object classification assume the same a priori occurrence probability for all considered object classes. However, with additional knowledge about the environment in which a scene was captured, the occurrence of some objects might be more likely than the occurrence of others. Taking into consideration this additional knowledge in the learning phase is called *context modelling*. In our approach the contexts are trained separately from the objects. For all considered contexts $\Upsilon_{\iota=1,\ldots,N_\Upsilon}$ the statistical context models $\mathcal{M}_{\iota=1,\ldots,N_\Upsilon}$ are learned. The context models contain a priori densities $p_\iota(\Omega_\kappa)$ for all objects classes $\Omega_{\kappa=1,\ldots,N_\Omega}$ taken into account in the recognition task. It is assumed that the number N_Υ and the types of context are known. Training starts with image acquisition where N_ι images are taken from random viewpoints with a hand-held camera for each context Υ_ι. The objects $\Omega_{\iota=1,\ldots,N_\Omega}$ occurring in the images are counted for each context. In the following $N_{\iota,\kappa}$ denotes how often the object Ω_κ occurs in the context Υ_ι. Generally, the sum of $N_{\iota,\kappa}$ for all object classes $\Omega_{\kappa=1,\ldots,N_\Omega}$ is not equal to N_ι. Therefore, for all contexts $\Upsilon_{\iota=1,\ldots,N_\Upsilon}$ a normalisation factor η_ι is defined as follows

$$\eta_\iota = \frac{N_\iota}{\sum_{\kappa=1}^{N_\Omega} N_{\iota,\kappa}}. \tag{14}$$

Using this normalisation factor η_ι and the number $N_{\iota,\kappa}$, the a priori occurrence probability for the object Ω_κ in the context Υ_ι is learned as

$$p_\iota(\Omega_\kappa) = \eta_\iota \frac{N_{\iota,\kappa}}{N_\iota}. \tag{15}$$

5 Classification and localisation

5.1 Single-object scenes

In this section it is assumed that each image contains exactly one single object. Moreover, the context modelling presented in Sect. 3.6 is not taken into consideration, i.e., the a priori probabilities for all possible object classes are assumed to be equal. In order to perform the classification and localisation in the image $\boldsymbol{f}$, the density values

$$p_{\kappa,h} = p(O_\kappa | \boldsymbol{B}_\kappa, \boldsymbol{\phi}_h, \boldsymbol{t}_h) \tag{16}$$

for all objects Ω_κ and for a large number of pose hypotheses $(\boldsymbol{\phi}_h, \boldsymbol{t}_h)$ are compared. First, the test image $\boldsymbol{f}$ is taken, preprocessed, and the local feature vectors $\boldsymbol{c}_m$ are determined according to Sect. 3.2. The computation of the object density value $p_{\kappa,h}$ for the given object Ω_κ, and pose parameters $(\boldsymbol{\phi}_h, \boldsymbol{t}_h)$ starts with the estimation of the object area $O_\kappa(\boldsymbol{\phi}_h, \boldsymbol{t}_h)$ which has been learned in the training phase (Sect. 3.3). For feature vectors from this object area $\boldsymbol{c}_m \in O_\kappa(\boldsymbol{\phi}_h, \boldsymbol{t}_h)$ the mean value vectors $\mu_{\kappa,m}$ and standard deviation vectors $\sigma_{\kappa,m}$ have been trained and are stored in the object models. Therefore, their density values

$$p_{\boldsymbol{c}_m} = p(\boldsymbol{c}_m | \boldsymbol{\mu}_{\kappa,m}, \boldsymbol{\sigma}_{\kappa,m}, \boldsymbol{\phi}_h, \boldsymbol{t}_h) \tag{17}$$

can be easily determined. Now, the object density value is calculated as follows

$$p_{\kappa,h} = \prod_{\boldsymbol{c}_m \in O_\kappa} \max\{p_{\boldsymbol{c}_m}, p_b\}, \tag{18}$$

where p_b is the background density introduced in Sect. 3.5. The reason for using the background density value p_b, if the density $p_{\boldsymbol{c}_m}$ is lower is explained in Fig. 3.

In case of an occlusion it might happen that the test feature vector $\boldsymbol{c}_m$ is completely different from the corresponding training feature represented by $\boldsymbol{\mu}_{\kappa,m}$ and $\boldsymbol{\sigma}_{\kappa,m}$. Thus, the density value for $\boldsymbol{c}_m$ would be very close to zero $p\left(\boldsymbol{c}_m | \boldsymbol{\mu}_{\kappa,m}, \boldsymbol{\sigma}_{\kappa,m}, \boldsymbol{\phi}_h, \boldsymbol{t}_h\right) \approx 0$. In this case the background density p_b is used in the product defined in (18). The object densities (18) normalised by a quality measure Q are maximised over all object classes Ω_κ and a large number of pose hypotheses $(\boldsymbol{\phi}_h, \boldsymbol{t}_h)$, as explained in Fig. 4. The quality measure (also called geometric criterion), defined in the following way

$$Q(p_{\kappa,h}) = \sqrt[N_{\kappa,h}]{p_{\kappa,h}} \tag{19}$$

decreases the influence the object size has on the recognition results. $N_{\kappa,h}$ denotes the number of feature vectors that belong to the object area $O_\kappa(\boldsymbol{\phi}_h, \boldsymbol{t}_h)$. The classification and localisation process presented in Fig. 4 can be described by the following maximisation term

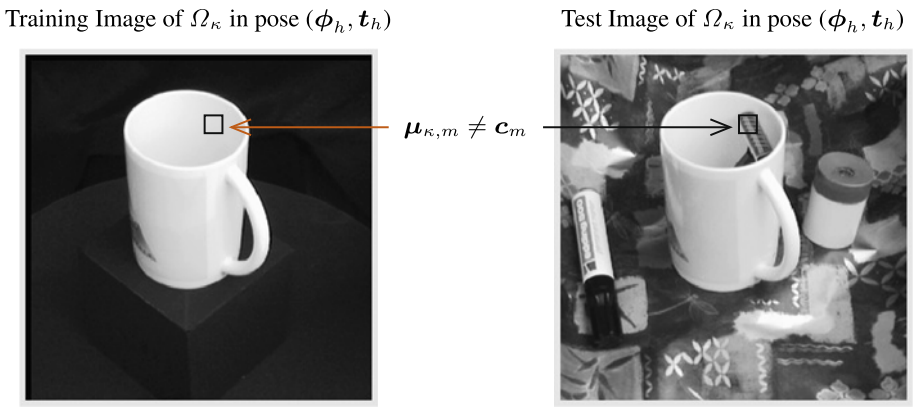

Fig. 3 Training image and test image of the same object in the same pose. Due to the occlusion with a razor in the test image, the test feature vector $\boldsymbol{c}_m$ is completely different from the corresponding training feature represented by $\boldsymbol{\mu}_{\kappa,m}$ and $\boldsymbol{\sigma}_{\kappa,m}$. Thus, the density value for $\boldsymbol{c}_m$ is very close to zero $p\left(\boldsymbol{c}_m|\boldsymbol{\mu}_{\kappa,m}, \boldsymbol{\sigma}_{\kappa,m}, \boldsymbol{\phi}_h, \boldsymbol{t}_h\right) \approx 0$.

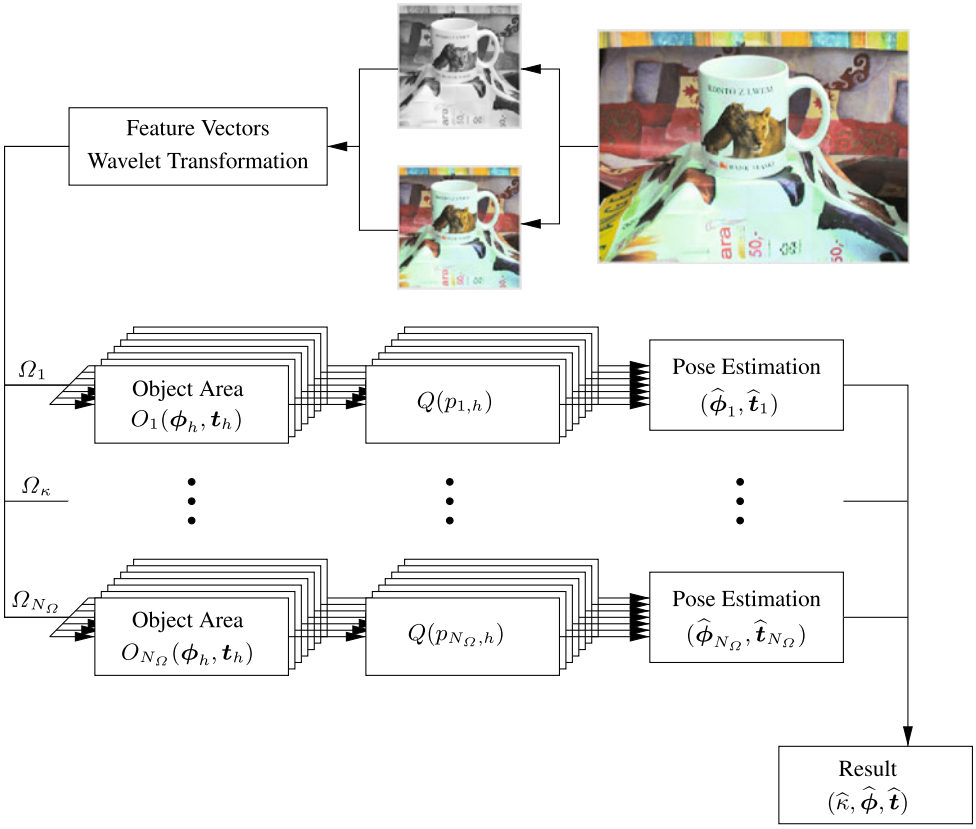

Fig. 4 Density maximisation for object classification and localisation. First, local feature vectors from the preprocessed test image are computed. Then, for each object class Ω_κ and each pose hypothesis $(\boldsymbol{\phi}_h, \boldsymbol{t}_h)$ the object area $O_\kappa(\boldsymbol{\phi}_h, \boldsymbol{t}_h)$ is determined and the object density $p_{\kappa,}h$ is calculated. The final recognition result $(\widehat{\kappa}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{t}})$ corresponds to the highest density normalised by a quality measure $Q(p_{\kappa,\ h})$

$$\left(\widehat{\kappa}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{t}}\right) = \underset{(\kappa,\boldsymbol{\phi}_h,\boldsymbol{t}_h)}{\operatorname{argmax}} Q(p(O_\kappa|\boldsymbol{B}_\kappa, \boldsymbol{\phi}_h, \boldsymbol{t}_h)) \tag{20}$$

where $(\widehat{\kappa}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{t}})$ represent the final recognition result, i.e., the class index and the pose parameters of the object found in image $\boldsymbol{f}$.

5.2 Multi-object scenes

In order to deal with multi-object scenes we make use of the modelled context dependencies (see Sect. 3.6). In the recognition phase there is no a priori knowledge about the context Υ_ι, in which the test image $\boldsymbol{f}$ has been taken. For this reason the algorithm automatically determines the context first. When searching for the first object Ω_{κ_1} in the multi-object scene $\boldsymbol{f}$, the algorithm does not make use of contextual information. The class k_1 and the pose $\left(\widehat{\boldsymbol{\phi}}_1, \widehat{\boldsymbol{t}}_1\right)$ of the first object is estimated by maximisation of the normalised object density value with (20), as illustrated in Fig. 4. It is assumed that at least one of the objects from the set $\Omega = \{\Omega_1, \Omega_2, \ldots, \Omega_\kappa, \ldots, \Omega_{N_\Omega}\}$ occurs in the image $\boldsymbol{f}$. Subsequently, the context $\Upsilon_{\widehat{\iota}}$ for the scene $\boldsymbol{f}$ (the context number $\widehat{\iota}$) is determined by maximisation of the a priori probability for the first object $p_{\iota=1,\ldots,N_\Upsilon}(\Omega_{\kappa_1})$ over all modelled contexts

$$\widehat{\imath} = \underset{\imath}{\operatorname{argmax}}\, p_{\imath}(\Omega_{\kappa_1}). \tag{21}$$

In the next step, the system estimates the optimal pose parameters $\left(\widehat{\phi}_{\kappa}, \widehat{\boldsymbol{t}}_{\kappa}\right)$ for all objects Ω_{κ} using the Maximum Likelihood (ML) method presented in Sect. 4.1:

$$\begin{aligned}
\left(\widehat{\phi}_1, \widehat{\boldsymbol{t}}_1\right) &= \underset{(\phi_h, \boldsymbol{t}_h)}{\operatorname{argmax}}\, Q(p(O_1|\boldsymbol{B}_1, \phi_h, \boldsymbol{t}_h)) \\
&\cdots \\
\left(\widehat{\phi}_{\kappa}, \widehat{\boldsymbol{t}}_{\kappa}\right) &= \underset{(\phi_h, \boldsymbol{t}_h)}{\operatorname{argmax}}\, Q(p(O_{\kappa}|\boldsymbol{B}_{\kappa}, \phi_h, \boldsymbol{t}_h)). \\
&\cdots \\
\left(\widehat{\phi}_{N_\Omega}, \widehat{\boldsymbol{t}}_{N_\Omega}\right) &= \underset{(\phi_h, \boldsymbol{t}_h)}{\operatorname{argmax}}\, Q(p(O_{N_\Omega}|\boldsymbol{B}_{N_\Omega}, \phi_h, \boldsymbol{t}_h))
\end{aligned} \tag{22}$$

Then, the object density values for the optimal pose parameters are weighted with the a priori probabilities $p_{\widehat{\imath}}(\Omega_{\kappa})$ learned for the context $\Upsilon_{\widehat{\imath}}$ in the training phase:

$$\begin{aligned}
\widehat{Q}_{\widehat{\imath},1} &= Q\left\{p_{\widehat{\imath}}(\Omega_1) p\left(O_1|\boldsymbol{B}_1, \widehat{\phi}_1, \widehat{\boldsymbol{t}}_1\right)\right\} \\
&\cdots \\
\widehat{Q}_{\widehat{\imath},\kappa} &= Q\left\{p_{\widehat{\imath}}(\Omega_{\kappa}) p\left(O_{\kappa}|\boldsymbol{B}_{\kappa}, \widehat{\phi}_{\kappa}, \widehat{\boldsymbol{t}}_{\kappa}\right)\right\} \\
&\cdots \\
\widehat{Q}_{\widehat{\imath},N_\Omega} &= Q\left\{p_{\widehat{\imath}}(\Omega_{N_\Omega}) p\left(O_{N_\Omega}|\boldsymbol{B}_{N_\Omega}, \widehat{\phi}_{N_\Omega}, \widehat{\boldsymbol{t}}_{N_\Omega}\right)\right\}
\end{aligned} \tag{23}$$

These normalised and weighted object densities $\widehat{Q}_{\widehat{\imath},\kappa=1,\ldots,N_\Omega}$ are then sorted in non-increasing order

$$\underbrace{\widehat{Q}_{\kappa_1} \geq \widehat{Q}_{\kappa_2}}_{d_1} \geq \cdots \geq \underbrace{\widehat{Q}_{\kappa_i} \geq \widehat{Q}_{\kappa_{i+1}}}_{d_i} \geq \cdots \geq \widehat{Q}_{\kappa_I}, \tag{24}$$

where $I = N_\Omega$ and d_i is a difference between neighbouring elements,

$$d_i = d(\widehat{Q}_{\kappa_i}, \widehat{Q}_{\kappa_{i+1}}) = \widehat{Q}_{\kappa_i} - \widehat{Q}_{\kappa_{i+1}} \tag{25}$$

The index $\widehat{i}$ of the highest distance $d_{\widehat{i}}$ $(\forall i \neq \widehat{i} : d_i \leq d_{\widehat{i}})$ is interpreted as the number of objects found in the multi-object scene $\boldsymbol{f}$ and is calculated as

$$\widehat{i} = \underset{i}{\operatorname{argmax}}\, d_i. \tag{26}$$

The final recognition results in the multi-object scene $\boldsymbol{f}$ are the following object classes and poses:

$$\begin{array}{ll}
\text{First object} & (\kappa_1, \widehat{\phi}_{\kappa_1}, \boldsymbol{t}_{\kappa_1}) \\
\text{Second object} & (\kappa_2, \widehat{\phi}_{\kappa_2}, \boldsymbol{t}_{\kappa_2}) \\
\vdots & \\
\text{Last object} & (\kappa_{\widehat{i}}, \widehat{\phi}_{\kappa_{\widehat{i}}}, \boldsymbol{t}_{\kappa_{\widehat{i}}})
\end{array} \tag{26}$$

Clearly, when evaluating recognition in scenes with multiple objects, not only the object classification result Ω_{κ_i} and the object localisation result $(\widehat{\phi}_{\kappa_i}, \boldsymbol{t}_{\kappa_i})$ have to be verified, but also the number $\widehat{i}$ of objects found in the scene $\boldsymbol{f}$.

6 Experiments and results

6.1 3D-REAL-ENV image database

In our experiments, we used the 3D-REAL-ENV [10, 11] database consisting of the ten real world objects depicted in Fig. 5. The pose of each object in the 3D-REAL-ENV

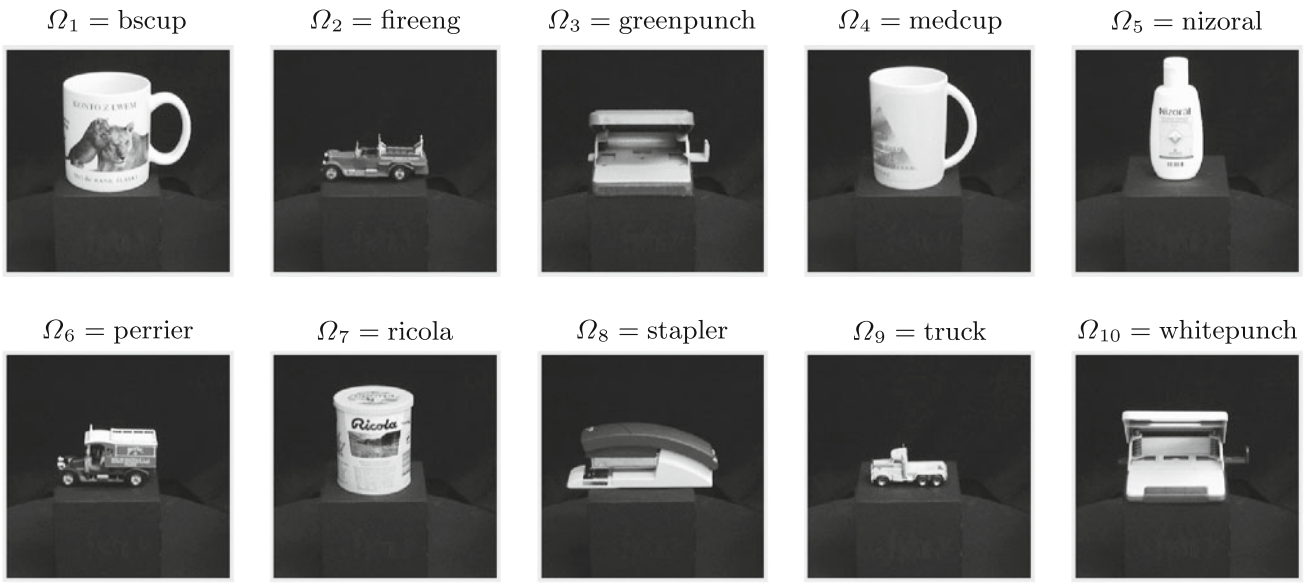

Fig. 5 Ten objects of the 3D-REAL-ENV image database with their short names. *Top row from left to right* bank cup, toy fire engine, green puncher, siemens cup, nizoral bottle. *Bottom row from left to right* toy passenger car, candy box, blue stapler, toy truck, white puncher

database is defined by internal translations $\boldsymbol{t}_{\text{int}} = (t_x, t_y)^{\text{T}}$ and external rotation parameters $\phi_{\text{ext}} = (\phi_x, \phi_y)^{\text{T}}$. The objects were captured in RGB at a resolution of 640 × 480 pixels under three different illumination settings $I_{\text{lum}} \in \{\text{bright}, \text{average}, \text{dark}\}$. For this experiment the captured images were resized to 256 × 256 pixels.

Training images were captured with the objects against a dark background from 1,680 different viewpoints under 2 different illumination settings $I_{\text{lum}} \in \{\text{bright,dark}\}$. This produced 3,360 training images in total for each 3D-REAL-ENV object. Each object was placed on a turntable performing a full rotation ($0° \leq \phi_{\text{table}} < 360°$) while the camera attached on a robotic arm was moved on a vertical to horizontal arc ($0° \leq \phi_{\text{arm}} \leq 90°$). The movement of the camera arm ϕ_{arm} corresponds to the first external rotation ϕ_x, while the turntable spin ϕ_{table} corresponds to the second external rotation parameter ϕ_y. The angle between two successive steps of the turntable corresponds to 4.5°. The rotation of the turntable induces an apparent translation in the object position in the image plane, which results in varying internal translation parameters $\boldsymbol{t}_{\text{int}} = (t_x, t_y)^{\text{T}}$. These translations parameters were determined manually after acquisition.

For testing, the 10 objects presented were captured from 288 different viewpoints under the average illumination setting ($I_{\text{lum}} = \text{average}$) and against 3 different backgrounds: homogeneous, weak heterogeneous, and strong heterogeneous. This resulted in three test sets of 2,880 images each denoted according to the background used as $T_{\text{ype}} \in \{\text{hom,weak,strong}\}$. Test scenes of the first type ($T_{\text{ype}} = \text{hom}$) were taken on homogeneous black background, while 200 different real backgrounds were used to create heterogeneous backgrounds ($T_{\text{ype}} \in \{\text{weak,strong}\}$). Examples of test images with all three types of background are shown in Fig. 6. Similarly to the acquisition of training images, the objects were put on a turntable ($0° \leq \phi_{\text{table}} < 360°$) and the camera moved on a robotic arm from vertical to horizontal ($0° \leq \phi_{\text{arm}} \leq 90°$). However, for test images the turntable's rotation between two successive steps is 11.25°, thus test views are generally different from the views used for training. Also, the illumination in the test scenes is different from the illumination in the training images.

6.2 Experimental results for single-object scenes

The recognition algorithm for single-object scenes described in Sect. 4.1 was evaluated for the 3D-REAL-ENV image database presented in the previous section. The training of statistical object models was performed for 6

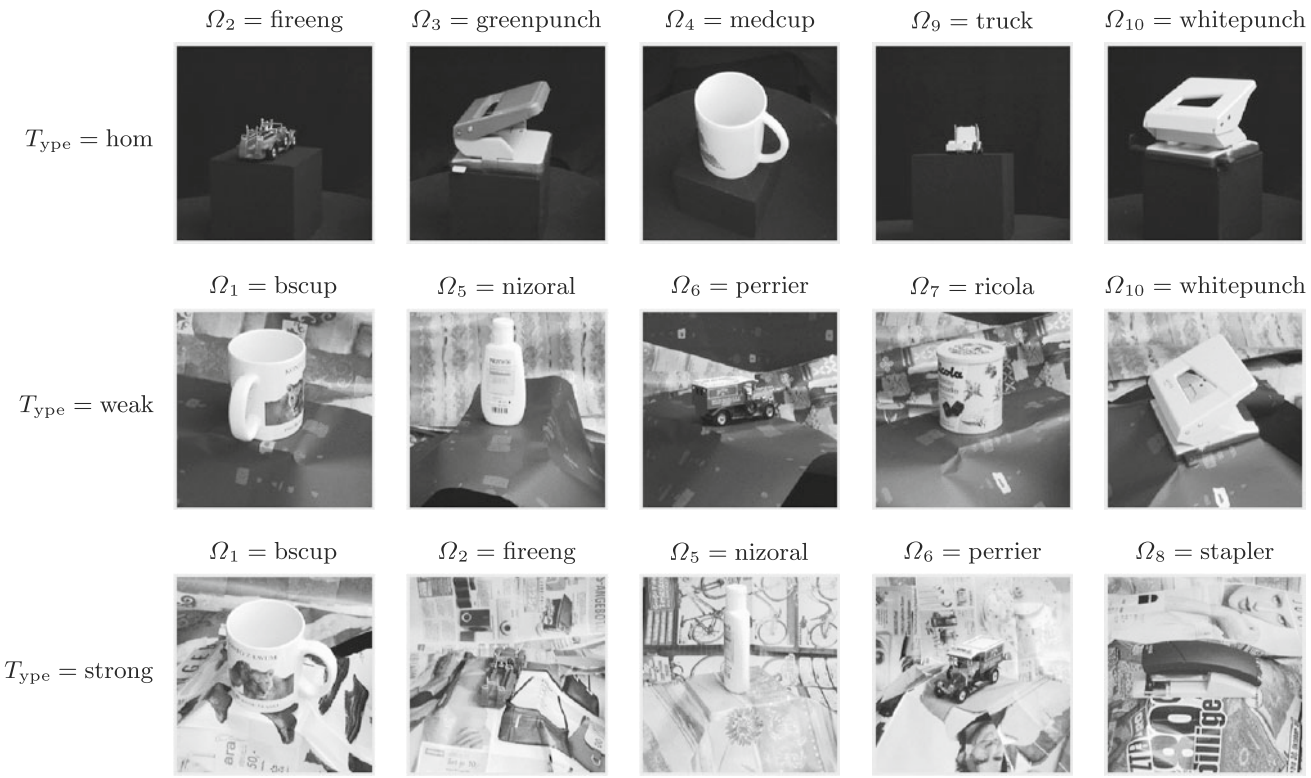

Fig. 6 Examples of test scenes on all three types of background $T_{\text{type}} \in \{\text{hom,weak,strong}\}$. The *top row* shows images with homogeneous background ($T_{\text{type}} = \text{hom}$), the *middle row* images with weak heterogeneous background ($T_{\text{type}} = \text{weak}$), and the *bottom row* images with strong heterogeneous background ($T_{\text{type}} = \text{strong}$)

Table 1 Classification and localisation rates obtained for 3D-REAL-ENV image database with gray-level and colour modelling

Distance of training views (°)	Type of object modelling	Classification rate (%)			Localisation rate (%)		
		Hom. Back.	Weak Het.	Strong Het.	Hom. Back.	Weak Het.	Strong Het.
4.5	Grey	100	92.2	54.1	99.1	80.9	69.0
	Colour	100	88.0	82.3	98.5	77.8	73.6
9.0	Grey	100	92.4	55.4	98.7	80.0	67.2
	Colour	100	88.3	81.2	98.2	76.4	72.1
13.5	Grey	99.4	89.7	56.2	96.9	78.6	65.4
	Colour	99.6	82.7	80.3	94.9	68.4	66.6
18.0	Grey	99.9	89.2	55.1	96.6	71.4	54.5
	Colour	97.3	80.6	68.6	94.3	64.9	60.7
22.5	Grey	99.4	86.0	52.8	94.5	60.7	38.6
	Colour	94.7	74.8	59.2	89.4	52.2	46.2
27.0	Grey	96.5	69.4	54.4	83.8	49.9	32.8
	Colour	93.8	53.6	50.2	78.3	35.8	35.6

The distance of training views varies from 4.5° to 27° in five steps. For experiments, 2,880 test images with homogeneous, 2,880 test images with weak heterogeneous, and 2,880 images with strong heterogeneous background were used

angle-steps (4.5°, 9°, 13.5°, 18°, 22.5°, 27°). Since this was done twice, i.e., for gray-level and colour images it resulted in 12 training configurations. The classification and localisation rates obtained for these configurations are summarised in Table 1. A classification result is counted as correct when the algorithm returns the correct object class. A localisation result is counted as correct when the error for internal translations is not greater than 10 pixels and the error for external rotations not greater than 15°. The results show that colour modelling brings a significant improvement in the classification and localisation rates for test images with strong heterogeneous background. For scenes with homogeneous background the recognition algorithm performs perfectly in both cases. For this type of background the computational expense associated with colour information can be avoided. However, in case of weak heterogeneous background better results for gray-level modelling are achieved. This is due to the assumption of statistical independency of RGB colour features which has been done in order to reduce the computation complexity. But it is obvious that the features computed from the red, the green, and the blue channel of a RGB image correlate with each other. In Table 1 one can also see how the recognition rates depend on the distance of training views. For test images with homogenous background, the classification algorithm works properly even for a high distance of 27°. However, it becomes worse in more complex scenarios. In case of colour modelling for test images with strong heterogeneous background the classification result is highly dependent on the distance of training views falling from 82.3% (4.5°) to 50.2% (27°). The localisation rates show their strong dependency on the distance of training views for all three test datasets. The reason for this is the necessary interpolation between the training views. Obviously, this interpolation works more precisely for denser data sets in terms of training viewpoints. Object recognition takes 3.6 s in one gray-level image and 7 s in one colour image on a workstation equipped with a Pentium 4, at 2.66 GHz, and 512 MB of RAM.

Table 2 Learned values of the a priori occurrence probabilities for all 3D-REAL-ENV object classes in three predefined contexts

	Ω_1	Ω_2	Ω_3	Ω_4	Ω_5	Ω_6	Ω_7	Ω_8	Ω_9	Ω_{10}
$\Upsilon 1$	0.20	0.06	0.06	0.20	0.04	0.06	0.20	0.06	0.06	0.06
$\Upsilon 2$	0.10	0.20	0.04	0.10	0.04	0.20	0.04	0.04	0.20	0.04
$\Upsilon 3$	0.10	0.04	0.20	0.10	0.04	0.04	0.04	0.20	0.04	0.20

6.3 Experimental results for multi-object scenes

For recognition of multi-object scenes, context modelling was incorporated in the system in addition to statistical object modelling. For each context considered in the experiments ($\Upsilon_1 = \text{kitchen}$, $\Upsilon_2 = \text{nursery}$, $\Upsilon_3 = \text{office}$), 100 images were captured with a hand-held camera at random viewpoints. Then, the a priori occurrence probabilities for all objects in all contexts (see Table 2) were trained as described in Sect. 3.6.

Altogether 3,240 grey-level multi-object scenes sized 512 × 512 pixels were used in the testing phase of the recognition algorithm. Each image contains between one and three objects from the 3D-REAL-ENV database pictured in Fig. 6. Similarly to the case of single-object scenes, the test images were divided into three types: 1,080

Table 3 Quantitative comparison of the system's performance with and without context modelling

3D-REAL-ENV image database	Without context modelling			With context modelling		
	Hom (%)	Weak (%)	Strong (%)	Hom (%)	Weak (%)	Strong (%)
ObjNumDet	100	83.9	43.2	99.9	88.2	59.2
Classification	100	91.9	62.9	100	97.0	87.5
Localisation	99.7	81.7	58.1	99.7	81.7	58.1

ObjNumDet Object number determination rate, *Hom* test images with homogeneous background, *Weak* test images with weak heterogeneous background, *Strong* test images with strong heterogeneous background

images with homogeneous background, 1,080 scenes with weak heterogeneous background, and 1,080 with strong heterogeneous background. Additionally, the 3D-REAL-ENV objects were assigned into three different contexts, namely the kitchen Υ_1, the nursery Υ_2, and the office Υ_3. For each background type and each context 120 one-object images, 120 two-object images, and 120 three-object images were created.

The quantitative comparison of our system's performance with and without context modelling is presented in Table 3. Since object localisation is performed for a priori known object classes, the context modelling does not influence its performance rate. However, the classification and the object number determination rates increases significantly when using context modelling for scenes with real heterogeneous background. This improvement is much greater for test images with strong heterogeneous background than for scenes with weak heterogenous background.

6.4 Experimental results for COIL image database

In order to allow a performance comparison of our system with other object recognition approaches, we performed additional experiments on the so called Columbia Object Image Library (COIL) image database. COIL-20 presented in [28] consists of 20 objects, while COIL-100 [27] is a completion of COIL-20 with additional 80 objects. Although the COIL image database provides only grey-level images and we could not make use of the colour modelling, we achieved satisfactory classification rates, namely 100% for COIL-20 and 98.9% for COIL-100.

6.5 Image quality impact on recognition results

Although the performance of the system for the real world dataset 3D-REAL-ENV is very good (see Table 1), the impact of the image quality on the recognition rates is quite high. Due to the statistical modelling it was possible to deal with illumination differences of about 20%[2]. However, considering illumination differences of about 50% the classification rate decreases from 82.3% for the test images with the strong heterogeneous background to 56% which cannot be accepted as a reasonable result anymore. Nevertheless, the system has proved an amazing robustness in terms of background heterogeneity. The strong heterogeneous background (see Fig. 6) can almost be considered as an attempt to camouflage the objects, and even for humans it turns out to be very difficult to find the objects in the images. In those extremely difficult conditions the system has still been able to achieve a classification rate of over 80%.

7 Real world application scenarios

7.1 Annotation of museum visit photos

A visit to a museum is an immersion into a rich universe of information not to mention an inspiring experience for young and old alike. However, it often happens that after spending a few hours in a museum we only remember some of the most impressive artefacts on display and even then only in general terms, particularly as the time elapsed since the visit increases. Fortunately, digital photo cameras are convenient extensions for our short-lived memory; pictures help us remember our experiences. Nowadays, cameras are omnipresent on holidays, excursions and cultural tours. Unfortunately, the sheer volume of data captured as a result typically prohibits useful manual indexing thereby detracting from the overall value of personal content collections. While traditionally cameras have been banned within museums, this is slowly changing. Museums may always need to restrict photography for specific exhibitions as a result of lender copyright agreements, but this becomes less critical when the museum owns the artefacts on display. As a result, increasingly museums are allowing photography and indeed are looking at new innovative uses of imaging technology that potentially afford a richer and longer lasting user experience. Of course, the "win–win" scenario that will drive real advances in this area is real benefit to both users and

[2] The difference between the mean grey level in the darkest and in the brightest image.

museums alike. In our proposed Web 2.0 service, the potential benefit to end users is richer annotation of their photos and access to relevant third-party premium content. For the museum, the potential benefit is a new source of revenue that leverages their existing cataloguing process in an innovative manner.

Our application is grounded in the stream of research that aims at bringing the benefits of digital technology for preservation, study and protection of heritage collections. Research initiatives such as SCULPTEUR[3] [8] and CHIP[4] [2] have targeted innovative ways of providing enhanced interaction and information content to museum visitors. Previously, radio frequency identification (RFID) tags have been used to guide visitors through discovery tours in museums and to provide enhanced information on the items of interest to visitors [21]. Image-based recognition of artefacts is potentially less expensive for museums as it can leverage the visitors own capture device, rather than putting in place an expensive RF tagging infrastructure and providing visitors with wireless PDA devices. Furthermore, it is not as encumbered by privacy concerns since the user movements are not being tracked directly.

In our application scenario, we envisage the museum offering a Web 2.0 service that automatically annotates a user's photos and offers added value content that can be personalised to the user based on their interest in specific artefacts. At some point after a visit, the user submits a set of digital photos taken inside the museum to the museum's site. The artefacts depicted are recognized using the approach presented in this paper providing a link between the users content and the museum's own data archives. The service then tags the user's photo with some metadata related to the artefact in question (e.g. name, date and location of origin, etc.). Drawing on their information rich catalogue, the museum can then offer relevant content to the user related to the artefact or exhibit he/she is interested in, such as access to online brochures, professional quality photos of the artefact itself or related artefacts that the user missed (note that typically only a small portion of the artefacts in a particular collection are used in an exhibit simply due to physical space restrictions), stylized renderings, 3D models, etc. While much of this valuable content exists in museum catalogues already as a result of a lengthy and expensive cataloguing process, it is difficult to provide access to it to the general public. In the proposed scenario the user may be willing to pay for access to these content sources enabling the museum to leverage additional revenue from its archiving activities. Figure 7 depicts a high-level conceptual diagram for a Web-based application that provides the above-mentioned service.

The crucial bottleneck in the proposed Web 2.0 service corresponds to automatic artefact identification, i.e., the classification process that should have the ability to accurately recognise the artefacts depicted in the submitted photos. The photos submitted by visitors are potentially quite diverse, being taken at various positions around the artefact on display. The scales at which the artefacts appear in various photos may also vary according to the distance to the camera and the zoom level used when the photo was captured. Therefore, the challenges in artefact recognition derive mainly from the changes in view (angle) and scale of the artefact in the photos. Clearly this is an ideal application scenario for the approach proposed in this paper. In order to deal with changes in position, multiple views of the artefact can be captured on a turntable that rotates the artefact in controlled steps around its own vertical axis during the museum's cataloguing process (very often a this happens anyway). Each photo to be recognised is then matched to multiple-views of artefacts in the collection captured in controlled conditions. Clearly, a multi-scale approach would be required to deal with scale variations and this is targeted as the next extension of our approach. We are currently designing and building a prototype application for this scenario in consultation with the National Museum of Ireland.

For preliminary experiments, we used an image database containing 75 artefacts. For training, 72 different viewpoints of all artefacts were used. For classification, 300 additional images under real museum-like conditions were acquired. Our system performed well for this image database and achieved a classification rate of 95.3%.

7.2 Classification of metallography images

The system presented in this contribution is being successfully applied for analysis of metallography images from the Ironworks in Ostrava (Czech Republic) [33]. The aim of this analysis is monitoring the quality process in the steel plant. Some example metallography images are presented in Fig. 8.

Metallography is a complex analysis process performed in the production of metal and composite materials with the purpose of controlling the composition and quality of the final alloy. This process involves various preparations of the metal specimen to be analysed followed by specialised visual inspection carried out under optical or electron microscopy. Based on the microscopy images a skilled technician can identify alloy composition and processing conditions. Considering the visual nature of the examination, metallography is an appealing test application for our texture-based image recognition approach.

[3] Semantic- and content-based multimedia exploitation for European benefit. http://www.sculpteurweb.org.

[4] Cultural Heritage Information Personalisation. http://www.chip-project.org.

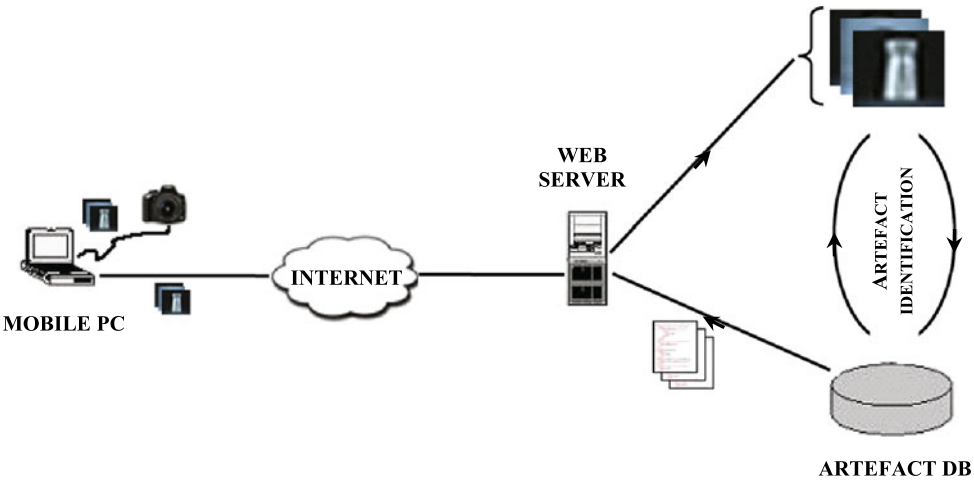

Fig. 7 Conceptual diagram of a Web-based artefact annotation system

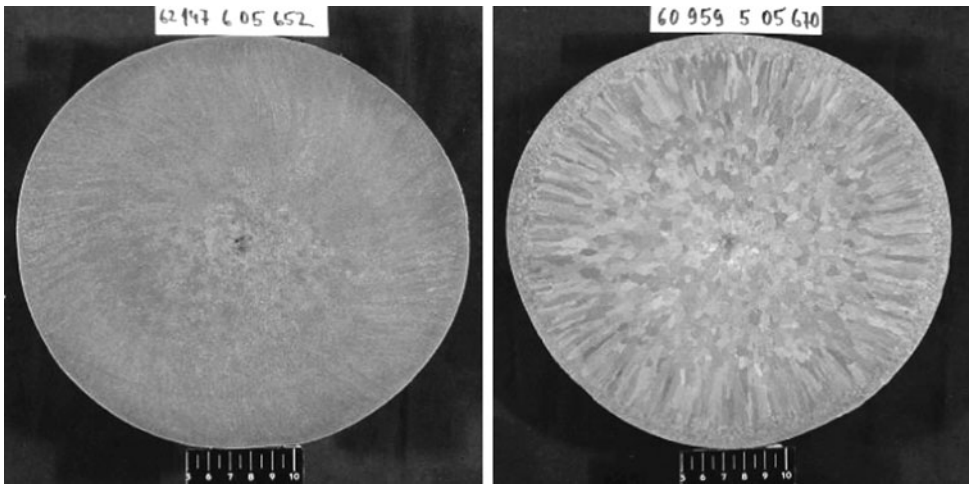

Fig. 8 Example metallography images from the Ironworks Ostrava (Czech Republic)

In order to classify metallography images into quality categories (image concepts) the object recognition problem reduces to an image classification task. The ground truth knowledge about the quality categories was provided by a human expert. The system has to find the concept $\Omega_{\widehat{\kappa}}$, (its index $\widehat{\kappa}$) present in a test image $\boldsymbol{f}$. For this, the density values for all concepts Ω_κ have to be compared to each other. Assuming the feature vectors $\boldsymbol{c}_m$ as statistically independent on each other the density value for the given test image $\boldsymbol{f}$ and concept Ω_κ is computed with

$$p_\kappa = \prod_{m=1}^{m=M} p(\boldsymbol{c}_m|\boldsymbol{\mu}_{\kappa,m}, \boldsymbol{\sigma}_{\kappa,m}), \tag{28}$$

where M is the number of all feature vectors in the image $\boldsymbol{f}$. All data required for computation of the density value p_κ with (28) are stored in the statistical concept model $\mathcal{M}_\kappa$. These density values are then maximised with ML estimation [40]

$$\widehat{\kappa} = \underset{\kappa}{\operatorname{argmax}}\, p_\kappa. \tag{29}$$

Having the index $\widehat{\kappa}$ of the resulting concept the classification problem for the image $\boldsymbol{f}$ is solved.

Preliminarily experiments carried out on this application show promising results [33]. We are continuing the work with a comprehensive investigation on quality scoring of metallography images, currently collecting data and setting up a large ground truth database.

8 Conclusions

This article presented a system for 3D texture-based probabilistic object classification and localisation in 2D images and discussed two real applications. The presented approach can be applied to both grey-level and colour images and to both single object and multi-object scenes whereby learned context dependencies between objects are leveraged in the latter case.

Experimental results on an image database of over 40,000 images illustrate the high performance of our system. A boost in performance is obtained using colour and context modelling. The classification rate achieved for 3D-REAL-ENV test images with strong heterogeneous background is 54.1% for grey-level modelling, while when

colour information is applied, the classification rate reaches 82.3%. The performance of the localisation algorithm is also improved by colour modelling for difficult heterogeneous environments, from 69.0% on grey-level modelling, to 73.6% on colour modelling. As a result of modelling of context dependencies between objects, higher classification rates were obtained for multi-object scenes. A classification rate of 62.9% is obtained for multi-object scenes with a strong heterogeneous background without considering context dependencies. Taking into account context increases the classification rate to 87.5%. The system described in this paper is currently being deployed in two quite different real application scenarios.

Improvements are possible with the approach and will form the basis of our future work. One line of research is to consider combining the appearance-based model with a shape-based model for object recognition. There are objects with the same shape, which are distinguishable only by texture, but one can also imagine objects with the same texture features, which can be easily distinguished by shape. Finally, since our system is adaptable to many image classification tasks we also intend to apply it in the context of knowledge assisted image and video content retrieval.

Acknowledgments This work was supported by the European Commission under the contract FP6-26978-X-MEDIA.

References

1. Amit Y, Geman D, Fan X (2004) A coarse-to-fine strategy for multi-class shape detection. IEEE Trans Pattern Anal Mach Intell 26(12):1606–1621
2. Aroyo L, Wang Y, Brussee R, Gorgels P, Rutledge L, Stash N (2007) Personalized museum experience: the Rijksmuseum use case. In: Proceedings of museums and the Web. San Francisco, USA
3. Bentoutou Y, Taleb N, Chikr El Mezouar M, Taleb M, Jetto L (2002) An invariant approach for image registration in digital subtraction angiography. Pattern Recognit 35(12):2853–2865
4. Chen H, Shimshoni I, Meer P (2004) Model based object recognition by robust information fusion. In: 17th international conference on pattern recognition. Cambridge, UK
5. Cho S-J, Kim JH (2004) Bayesian network modeling of strokes and their relationships for on-line handwriting recognition. Pattern Recognit 37(2):253–264
6. Fründ J, Gausemeier J, Matysczok C, Radkowski R (2004) Using augmented reality technology to support the automobile development. In: Shen W, Lin Z, Barthès J-PA, Li T (eds) 8th international conference on computer supported cooperative work in design. Springer, Xiamen, pp 289–298
7. Gausemeier J, Grafe M, Matysczok C, Radkowski R, Krebs J, Oelschlaeger H (2005) Eine mobile augmented reality versuchsplattform zur untersuchung und evaluation von fahrzeuggergonomien. In: Schulze T, Horton G, Preim B, Schlechtweg S (eds) Simulation und Visualisierung. SCS Publishing House e.V., Magdeburg, pp 185–194
8. Goodall S, Lewis PH, Matrinez K, Sinclair PAS, Giorgini F, Addis MJ, Boniface MJ, Lahanier C, Stevenson J (2004) Sculpteur: multimedia retrieval for museums. In: Third international conference on image and video retrieval (CIVR 2004). Dublin, Ireland, pp 638–646
9. Gross R, Matthews I, Baker S (2004) Appearance-based face recognition and light-fields. IEEE Trans Pattern Anal Mach Intell 26(4):449–465
10. Grzegorzek M, Niemann H (2005) Statistical object recognition including color modeling. In: Kamel M, Campilho A (eds) 2nd international conference on image analysis and recognition. Lecture Notes in Computer Science, vol 3656. Toronto, Canada, Springer, Berlin, pp 481–489
11. Grzegorzek M, Reinhold M, Niemann H (2005) Feature extraction with wavelet transformation for statistical object recognition. In: Kurzynski M, Puchala E, Wozniak M, Zolnierek A (eds) 4th international conference on computer recognition systems. Rydzyna, Poland, Springer, Berlin, pp 161–168
12. Heigl B (2004) Plenoptic scene modeling from uncalibrated image sequences. Ibidem-Verlag, Stuttgart
13. Heutte L, Nosary A, Paquet T (2004) A multiple agent architecture for handwritten text recognition. Pattern Recognit 37(4):665–674
14. Hornegger J (1996) Statistische Modellierung, Klassifikation und Lokalisation von Objekten. Shaker Verlag, Aachen
15. Jin Y, Geman S (2006) Context and hierarchy in a probabilistic image model. In : IEEE conference on computer vision and pattern recognition. New York, USA, pp 2145–2152
16. Kerr J, Compton P (2003) Toward generic model-based object recognition by knowledge acquisition and machine learning. In: Proceedings of the eighteenth international joint conference on artificial intelligence. Acapulco, Mexico, pp 9–15
17. Kumar A (2003) Neural network based detection of local textile defects. Pattern Recognit 36(7):1631–1644
18. Latecki LJ, Lakaemper R, Wolter D (2005) Optimal partial shape similarity. Image Vis Comput J 23:227–236
19. Leibe B, Schiele B (2003) Analyzing contour and appearance based methods for object categorization. In: IEEE conference on computer vision and pattern recognition. Madison, USA
20. Li CH, Yuen PC (2002) Tongue image matching using color content. Pattern Recognit 35(2):407–419
21. Liu TY, Tan TH, Chu YL (2006) The ubiquitous museum learning environment: concept, design, implementation, and a case study. In: Sixth international conference on advanced learning technologies. Kerkrade, The Netherlands, pp 989–991
22. Lowe DG (1999) Object recognition from local scale-invariant fearures. In: Seventh international conference on computer vision (ICCV). Corfu, Greece, pp 1150–1157
23. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
24. Mahamud S, Hebert M (2003) The optimal distance measure for object detection. In: IEEE conference on computer vision and pattern recognition. Madison, USA
25. Mallat S (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 11(7):674–693
26. Murase H, Nayar SK (1995) Visual learning and recognition of 3-D objects from appearance. Int J Comput Vis 14(1):5–24
27. Nene S, Nayar S, Murase H (1996) Columbia object image library (coil-100). Technical Report CUCS-006-96, Department for Computer Science, Columbia University
28. Nene S, Nayar S, Murase H (1996) Columbia object image library (coil-20). Technical Report CUCS-005-96, Department for Computer Science, Columbia University
29. Ngan HYT, Pang GKH, Yung SP, Ng MK (2005) Wavelet based methods on patterned fabric defect detection. Pattern Recognit 38(4):559–576

30. Park CH, Park H (2005) Fingerprint classification using fast fourier transform and nonlinear discriminant analysis. Pattern Recognit 38(4):495–503
31. Pontil M, Verri A (1998) Support vector machines for 3D object recognition. IEEE Trans Pattern Anal Mach Intell 20(6):637–646
32. Pösl J (1999) Erscheinungsbasierte, statistische Objekterkennung. Shaker Verlag, Aachen
33. Praks P, Grzegorzek M, Moravec R, Valek L, Izquierdo E (2007) Wavelet and eigen-space feature extraction for classification of metallography images. In: Jaakkola H, Kiyoki Y, Tokuda T (eds) European–Japanese conference on information modeling and knowledge bases. Juvenes Print-TTY, Tampere, Pori, Finland, pp 193–202
34. Reinhold M (2004) Robuste, probabilistische, erscheinungsbasierte Objekterkennung. Logos Verlag, Berlin
35. Rue H, Held L (2005) Gaussian Markov random fields: theory and applications. Chapman & Hall, London
36. Schiele B, Crowley JL (2000) Recognition without correspondence using multidimensional receptive field histograms. Int J Comput Vis 36(1):31–50
37. Schneiderman H, Kanade T (2004) Object detection using the statistics of parts. Int J Comput Vis 56(3):151–177
38. Terzopoulos D, Yuencheng L, Vasilescu M (2004) Model-based and image-based methods for facial image synthesis, analysis and recognition. In: Automatic face and gesture recognition 2004. Seoul, Korea, pp 3–8
39. Torralba A, Murphy KP, Freeman WT (2007) Sharing visual features for multiclass and multiview object detection. IEEE Trans Pattern Anal Mach Intell 29(5):854–869
40. Webb AR (2002) Statistical pattern recognition. Wiley, Chichester
41. You B, Hwangbo M, Lee S, Oh S, Kwon Y, Lim S (2003) Development of a home service robot issac. In: Intelligent robots and systems 2003, Las Vegas, USA, pp 2630–2635
42. Zhang Q, Yan H (2004) Fingerprint classification based on extraction and analysis of singularities and pseudo ridges. Pattern Recognit 37(11):2233–2243
43. Zobel M, Denzler J, Heigl B, Nöth E, Paulus D, Schmidt J, Stemmer G (2003) Mobsy: integration of vision and dialogue in service robots. Mach Vis Appl 14(1):26–34

Author Biography

Marcin Grzegorzek Marcin Grzegorzek, born in 1977, graduated in Computer Science at the Silesian University of Technology in Gliwice (Poland) in 2002. From 2002 to 2006 he worked for the Institute of Pattern Recognition at the University of Erlangen-Nuremberg (Germany), where he received his PhD with distinction in the field of 3D statistical object classification and localisation. Then he moved to the Multimedia and Vision Research Group at the Queen Mary, University of London (UK) and worked on multimedia analysis and retrieval.2008 he joined the University of Koblenz-Landau(Germany) and works on semantically driven image analysis and cross-media technologies. He leads a Focus Group for Multimedia Web (MMWeb) providing a scientific bridge between several Research Groups in the Department of Computer Science. Dr. Grzegorzek presented his scientific results on international conferences and workshops, in international journals, as well as in form of a text book. He successfully worked on and managed national and international projects. He has participated in several conference program and organising committees. He is a Guest Editor of the International Journal on Multimedia Tools and Applications. Moreover, he acts as a Secretary in the Executive Board of the SMaRT (Semantic Multimedia Research and Technology) Association which aims at offering an integrative scientific platform for leading researchers from the fields of Semantic Web, Multimedia and Signal Analysis. Finally, he is a Founding Committee Member of the Institute for Medicine Engineering and Information Processing (MTI Mittelrhein), an interdisciplinary facility bringing together researchers from the Mittelrhein area who work on new solutions for medical applications.