Local Wavelet Features for Statistical Object Classification and Localisation

Marcin Grzegorzek*
ISWeb – Information Systems and Semantic Web Resaerch Group
Institute for Computer Science, University of Koblenz – Landau
Universitätsstraße 1, 56070 Koblenz, Germany
Phone: +49-261-287-1251
Fax: +49-261-287-2721
marcin@uni-koblenz.de

Sorin Sav
Centre for Digital Video Processing
Dublin City University
Glasnevin, Dublin 9, Ireland
Phone: +35-31-700-6830
Fax: +35-31-700-5508
sorinsav@eeng.dcu.ie

Ebroul Izquierdo, *Senior Member, IEEE*
Head of the Multimedia & Vision Research Group
Queen Mary, University of London
Mile End Road, E1 4NS London, UK
Phone: +44-20-7882-5354
Fax: +44-20-7882-7997
ebroul.izquierdo@elec.qmul.ac.uk

Noel E. O'Connor, *Member, IEEE*
Centre for Digital Video Processing
Dublin City University
Glasnevin, Dublin 9, Ireland
Phone: +35-31-700-5078
Fax: +35-31-700-5508
oconnorn@eeng.dcu.ie

***Abstract*—This article presents a system for texture-based probabilistic classification and localisation of 3D objects in 2D digital images and discusses selected applications. The objects are described by local feature vectors computed using the wavelet transform. In the training phase, object features are statistically modelled as normal density functions. In the recognition phase, a maximisation algorithm compares the learned density functions with the feature vectors extracted from a real scene and yields the classes and poses of objects found in it. Experiments carried out on a real dataset of over 40000 images demonstrate the robustness of the system in terms of classification and localisation accuracy. Finally, two important application scenarios are discussed, namely classification of museum artefacts and classification of metallography images.**

***Index Terms*—Object Recognition, Statistical Modelling, Wavelet Analysis, Image Processing**

I. Introduction

A fundamental problem of *computer vision* is the recognition of objects in digital images. The term *object recognition* covers both, *classification* and *localisation* of objects. For the problem of object classification the system must 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 symmetry centre of the image, the x- and y-axes lie in the image plane, and the z-axis is orthographic to the image plane. These transformation parameters are divided into internal ($\boldsymbol{t}_{\mathrm{int}} = (t_x, t_y)^{\mathrm{T}}$, $\phi_{\mathrm{int}} = \phi_z$) for 2D objects and external ($t_{\mathrm{ext}} = t_z$, $\boldsymbol{\phi}_{\mathrm{ext}} = (\phi_x, \phi_y)^{\mathrm{T}}$) for 3D objects.

For recognition of 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 as well as relations between them are used for object description [1]. 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 using the image data, i. e., the pixel values, directly without a previous segmentation step. For this reason the texture-based method for object recognition has been chosen to develop the system presented in this contribution.

The object recognition problem has been intensively investigated in the past. Many approaches to object recognition, like the one presented in this paper, are founded on probability theory [2], and can be broadly characterised as either generative or discriminative according to whether or not the distribution of the image features is modelled [3]. Generative models such as principal component analysis (PCA) [4], independent component analysis (ICA) [5] or non-negative matrix factorisation (NMF) [6] try to find a suitable representation of the original data [7]. In contrast, discriminative classifiers such as linear discriminant analysis (LDA) [8], support vector machines (SVM) [9], [10], or boosting [11] aim at finding optimal decision boundaries given the training data and the corresponding labels [7]. The system presented in this paper represents the generative approaches.

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 [12], [13], fingerprint classification [14], handwriting recognition [15], service robotics [16], medicine [17], visual inspection [18], the automobile industry [13], [19], 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 classification of artefacts following a visit to a museum, another is the analysis of metallography images from an ironworks.

There are further interesting approaches for object recognition. Amit et al. proposes in [20] an algorithm for multi-class shape detection in the sense of recognising and localising instances from multiple shape classes. In [21] 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. In [22] the problem of detecting a large number of different classes of objects in cluttered scenes is taken into consideration. [23] proposes a mathematical framework for constructing probabilistic hierarchical image models, designed to accommodate arbitrary contextual relationships. In order to compare different methods for object recognition, in [24] a new database specifically tailored to the task of object categorisation is presented. In [25] 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. In [26] a multi-class object detection framework whose core component is a nearest neighbour search over object part classes is presented.

As can be seen above, a lot of valuable research work has been done in the field of object recognition in the past. However, many features of our system prove its novelty and originality as well as high performance in the sense of classification and localisation accuracy. One of them is the fusion of multiple views based on a recursive density propagation. Furthermore, the training phase in our system can be performed using images taken with a hand-held camera. The missing pose parameters are then automatically reconstructed with the so called structure-from-motion algorithm [27]. In order to improve the performance of our system, we also introduced the colour and the context modelling. Moreover, the object feature extraction can be performed on different resolution levels of the wavelet transform [28]. The object models learned for these different resolutions can be then combined with each other to accelerate the search and improve the recognition results.

Many of the system features are presented on the following pages. Section II describes the training procedure for the object and context modelling. The object recognition phase is detailed in Section III. Section IV covers the experimental results achieved on a large database of over 40000 images of real objects captured against heterogeneous backgrounds. Section V describes two real application scenarios successfully implemented with our system: classification of museum artefacts and classification of metallography images. The final conclusions of this work are presented in Section VI.

II. Training

This section starts with a short description of the acquisition of data for training in Section II-A, followed by an explanation on the feature extraction process in Section II-B. The so called "object area" is then defined in Section II-C. The statistical methods for object and background modelling are presented in Sections II-D and II-E respectively. Finally, Section II-F briefly presents the statistical context modelling, which can also be performed using the system in training mode. Since the training process is identical for all objects Ω_κ, the object class index κ will be omitted ($\Omega_\kappa = \Omega$) until the end of Section II-E.

A. 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 their external pose parameters $(\boldsymbol{\phi}_{\mathrm{ext}}, t_{\mathrm{ext}})$ known for each training image.

B. Feature Extraction with 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 $\boldsymbol{c}_m$ in these images are computed via the discrete wavelet transform [28]. In order to calculate the $\boldsymbol{c}_m$ vectors, a grid with the size $\Delta r = 2^{|\widehat{s}|}$, where $\widehat{s}$ is the minimum multiresolution scale parameter[1] s, is overlaid on the image [29]. Figure 1 depicts this procedure for the case of gray level scenes divided into local neighbourhoods of size 4×4 pixels. Using the coefficients introduced in Figure 1, the local feature vector $\boldsymbol{c}_m$ for the gray level image is defined by,

$$\boldsymbol{c}_m = \begin{pmatrix} \ln(2^{\widehat{s}}|b_{\widehat{s}}|) \\ \ln[2^{\widehat{s}}(|d_{0,\widehat{s}}| + |d_{1,\widehat{s}}| + |d_{2,\widehat{s}}|)] \end{pmatrix} \quad . \tag{1}$$

In the feature vector, the first component stores information about the mean gray level (low-frequencies) in the local neighbourhood, while the second component represents discontinuities (high-frequencies). The natural logarithm (ln) helps to depress local artefacts which can occur in real environments. 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 Figure 1). Therefore, the local feature vector for colour images has six components. 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

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

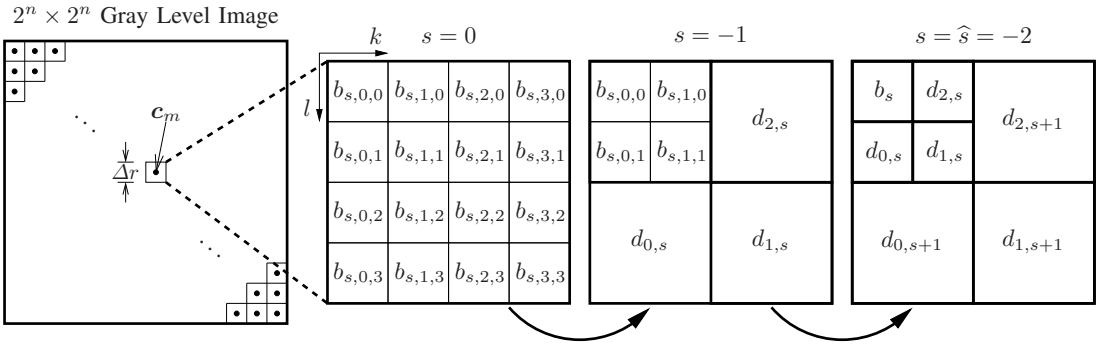

Figure 1. 2D signal decomposition with the wavelet transform for a local neighbourhood of size 4×4 pixels. The final coefficients result from gray values $b_{0,k,l}$ and have the following meaning: b_{-2} : low-pass horizontal and low-pass vertical, $d_{0,-2}$: low-pass horizontal and high-pass vertical,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

the sixth $c_{m,6}$ from the blue channel. 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.

C. Object Area Definition

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*. By this term the set of features belonging to the object will be referred. 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 $\phi_{\text{int}} \in \mathbb{R}$ around the perpendicular axis to the image plane and translated by $\boldsymbol{t}_{\text{int}} \in \mathbb{R}^2$ in the image plane, its appearance 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 $\boldsymbol{c}_m$ describes the object ($\boldsymbol{c}_m \in O$), whilst for others the same vector belongs to the background ($\boldsymbol{c}_m \notin O$). For this reason, the object area is modelled as a function of the external pose parameters

$$O = O(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}) \quad , \tag{2}$$

ideally within a continuous domain. This is done by 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,

$$\xi = \xi_m(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}) \quad . \tag{3}$$

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,

$$\begin{cases} \xi_m(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}) \geq S_O & \Rightarrow \quad \boldsymbol{c}_m \in O(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}) \\ \xi_m(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}) < S_O & \Rightarrow \quad \boldsymbol{c}_m \notin O(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}) \end{cases} , \tag{4}$$

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

$$\xi_m(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}) = \begin{cases} 1, & \text{if} \quad c_{m,1} \geq S_\xi \\ 0, & \text{if} \quad c_{m,1} < S_\xi \end{cases} , \tag{5}$$

where S_ξ is a threshold value[2]. 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

$$O = O(\boldsymbol{\phi}, \boldsymbol{t}) \tag{6}$$

defined in a continuous six-dimensional pose parameter space $(\boldsymbol{\phi}, \boldsymbol{t})$. Please note that an object feature vector ($\boldsymbol{c}_m \in O$) for a particular view $(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}})$ is always computed on a particular object point $\boldsymbol{x}_m$, i. e., it moves with the object within the image plane in terms of internal pose parameters $(\phi_{\text{int}}, \boldsymbol{t}_{\text{int}})$.

D. Statistical Object Modelling

In order to handle illumination changes and low-frequency noise, the elements $c_{m,q}$ of the local feature vectors $\boldsymbol{c}_m$ are interpreted as normal 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 Section II-E. The elements of the object feature vectors are represented with Gaussian density functions $p(c_{m,q}|\mu_{m,q}, \sigma_{m,q}, \boldsymbol{\phi}, \boldsymbol{t})$. The mean $\mu_{m,q}$ and standard deviation $\sigma_{m,q}$ values are estimated for all training views $(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}})$, which form a subspace of $(\boldsymbol{\phi}, \boldsymbol{t})$. Assuming the statistical independence of the elements $c_{m,q}$, which is valid due to their different interpretations in terms of signal processing (Section II-B), the density function for the object feature vector $\boldsymbol{c}_m \in O$ can be written as,

$$p(\boldsymbol{c}_m|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m, \boldsymbol{\phi}, \boldsymbol{t}) = \prod_{q=1}^{N_q} p(c_{m,q}|\mu_{m,q}, \sigma_{m,q}, \boldsymbol{\phi}, \boldsymbol{t}) \quad , \tag{7}$$

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 supposed that the feature vectors belonging to the object $\boldsymbol{c}_m \in O$ are statistically independent of each other. Under

[2]In the training phase objects are acquired against homogeneous background, either black (bright objects) or white (dark objects). Therefore, a simple thresholding is sufficient for object area detection. (5) assumes bright objects and would change its direction for dark ones.

this assumption, an object can be described by the probability density p as follows,

$$p(O|\boldsymbol{B},\boldsymbol{\phi},\boldsymbol{t}) = \prod_{\boldsymbol{c}_m \in O} p(\boldsymbol{c}_m|\boldsymbol{\mu}_m,\boldsymbol{\sigma}_m,\boldsymbol{\phi},\boldsymbol{t}) \quad , \tag{8}$$

where $\boldsymbol{B}$ comprehends the mean value vectors $\boldsymbol{\mu}_m$ and the standard deviation vectors $\boldsymbol{\sigma}_m$. This probability density is called the *object density* and, taking into account (7), can be written in more detail as,

$$p(O|\boldsymbol{B},\boldsymbol{\phi},\boldsymbol{t}) = \prod_{\boldsymbol{c}_m \in O} \prod_{q=1}^{N_q} p(c_{m,q}|\mu_{m,q},\sigma_{m,q},\boldsymbol{\phi},\boldsymbol{t}) \quad . \tag{9}$$

In order to complete the object description with the object density (9), 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 (9) over all N_ρ training images,

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\sigma}}) = \underset{(\boldsymbol{\mu},\boldsymbol{\sigma})}{\operatorname{argmax}} \prod_{\rho=1}^{N_\rho} p(O|\boldsymbol{B},\boldsymbol{\phi}_\rho,\boldsymbol{t}_\rho) \quad . \tag{10}$$

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.

E. Statistical Background Modelling

As mentioned in Section II-D, 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 $c_{m,q}$ can be observed with the same probability. Thus, they are modelled as uniform random variables, and their constant density functions,

$$p(c_{m,q}) = \frac{1}{\max(c_{m,q}) - \min(c_{m,q})} \tag{11}$$

do not depend on the transformation parameters $(\boldsymbol{\phi}, \boldsymbol{t})$. Assuming the statistical independence of $c_{m,q}$, (11) can be extended to

$$p(\boldsymbol{c}_m) = \prod_{q=1}^{N_q} \frac{1}{\max(c_{m,q}) - \min(c_{m,q})} = p_b \quad , \tag{12}$$

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

F. 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. The training starts with the image acquisition where N_ι images are taken from random viewpoints with a hand-held camera for each context Υ_ι. The objects $\Omega_{\kappa=1,\ldots,N_\Omega}$ 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. This number defines the a-priori occurrence probability for the object Ω_κ in the context Υ_ι as follows

$$p_\iota(\Omega_\kappa) = \eta_\iota N_{\iota,\kappa} \quad , \tag{13}$$

whereas the normalisation factor η_ι ensures that the sum of the a-priori occurrence probabilities for all objects in the context Υ_ι is equal to 1.

III. Classification and Localisation

This section describes the recognition mode of the system. The classification and localisation algorithm for single-object scenes is presented in Section III-A, while Section III-B deals with multi-object scenes.

A. Single-Object Scenes

In this section it is assumed that each image contains exactly one single object. In order to perform the classification and localisation in the image $\boldsymbol{f}$, the density values

$$p_{\kappa,h} = p(O_\kappa|\boldsymbol{B}_\kappa,\boldsymbol{\phi}_h,\boldsymbol{t}_h) \tag{14}$$

for all objects Ω_κ and for a large number of pose hypotheses $(\boldsymbol{\phi}_h, \boldsymbol{t}_h)$ are compared to each other. As you can see, the pose parameter space has been discretised again $(\boldsymbol{\phi}_h, \boldsymbol{t}_h)$ and the training interpolation to a fully continuous model (see Section II-D) might seem to had been unnecessary. However, the time optimisation in the recognition phase has got a higher priority than the time reduction in the training process. First, the test image $\boldsymbol{f}$ is taken, preprocessed, and the local feature vectors $\boldsymbol{c}_m$ are determined according to Section II-B. 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 (Section II-C). For feature vectors from this object area $\boldsymbol{c}_m \in O_\kappa(\boldsymbol{\phi}_h, \boldsymbol{t}_h)$ the mean value vectors $\boldsymbol{\mu}_{\kappa,m}$ and standard deviation vectors $\boldsymbol{\sigma}_{\kappa,m}$ have been trained and are stored in the object models. Therefore, their density values

$$p_{\boldsymbol{c}_m} = p(\boldsymbol{c}_m|\boldsymbol{\mu}_{\kappa,m},\boldsymbol{\sigma}_{\kappa,m},\boldsymbol{\phi}_h,\boldsymbol{t}_h) \tag{15}$$

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

$$p_{\kappa,h} = \prod_{\boldsymbol{c}_m \in O_\kappa} \max\{p_{\boldsymbol{c}_m}, p_b\} \quad , \tag{16}$$

where p_b is the background density introduced in Section II-E. This is applied as a minimum multiplication component in

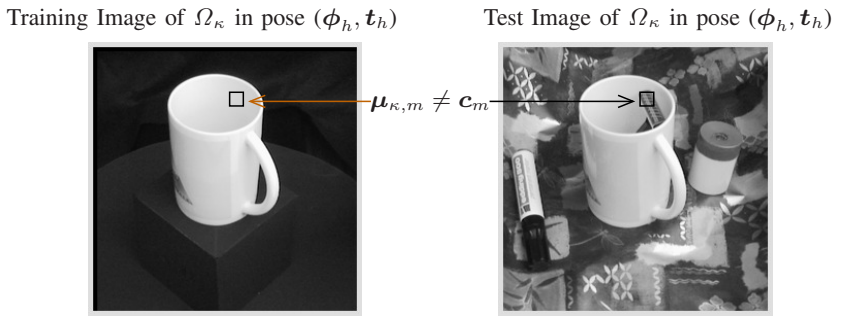

Figure 2. Training image and test image of the same object in the same pose. Due to the occlusion with a razor in the test image, the test feature vector $\boldsymbol{c}_m$ is completely different from the corresponding training feature represented by $\boldsymbol{\mu}_{\kappa,m}$ and $\boldsymbol{\sigma}_{\kappa,m}$. Thus, the density value for $\boldsymbol{c}_m$ is very close to zero $p(\boldsymbol{c}_m|\boldsymbol{\mu}_{\kappa,m}, \boldsymbol{\sigma}_{\kappa,m}, \boldsymbol{\phi}_h, \boldsymbol{t}_h) \approx 0$.

order to solve object occlusions such as that presented in Figure 2. The object densities (16) normalised by a quality measure Q are maximised over all object classes Ω_κ and a large number of pose hypotheses. The quality measure (also called geometric criterion), defined in the following way

$$Q(p_{\kappa,h}) = \sqrt[N_{\kappa,h}]{p_{\kappa,h}} \quad , \tag{17}$$

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 can be described by the following maximisation term

$$(\widehat{\kappa}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{t}}) = \underset{(\kappa,\boldsymbol{\phi}_h,\boldsymbol{t}_h)}{\operatorname{argmax}} Q(p(O_\kappa|\boldsymbol{B}_\kappa, \boldsymbol{\phi}_h, \boldsymbol{t}_h)) \tag{18}$$

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 $\boldsymbol{f}$.

B. Multi-Object Scenes

This section deals with multi-object scenes under consideration of context dependencies. These context dependencies have been modelled in the training phase as described in Section II-F. 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 κ_1 and the pose $(\widehat{\boldsymbol{\phi}}_1, \widehat{\boldsymbol{t}}_1)$ of the first object is estimated by maximisation of the normalised object density value with (18). 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

$$\widehat{\iota} = \underset{\iota}{\operatorname{argmax}}\, p_\iota(\Omega_{\kappa_1}) \quad . \tag{19}$$

In the next step, the system estimates the optimal pose parameters $(\widehat{\boldsymbol{\phi}}_\kappa, \widehat{\boldsymbol{t}}_\kappa)$ for all objects $\Omega_{\kappa=1,\ldots,N_\Omega}$ using the Maximum Likelihood (ML) method presented in Section III-A

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

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

$$\widehat{Q}_{\widehat{\iota},\kappa} = Q\{p_{\widehat{\iota}}(\Omega_\kappa) p(O_\kappa|\boldsymbol{B}_\kappa, \widehat{\boldsymbol{\phi}}_\kappa, \widehat{\boldsymbol{t}}_\kappa) \quad . \tag{21}$$

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

$$\underbrace{\widehat{Q}_{\kappa_1} \geq \widehat{Q}_{\kappa_2}}_{d_1} \geq \ldots \geq \underbrace{\widehat{Q}_{\kappa_i} \geq \widehat{Q}_{\kappa_{i+1}}}_{d_i} \geq \ldots \geq \widehat{Q}_{\kappa_I} \quad , \tag{22}$$

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

$$d_i = d(\widehat{Q}_{\kappa_i}, \widehat{Q}_{\kappa_{i+1}}) = \widehat{Q}_{\kappa_i} - \widehat{Q}_{\kappa_{i+1}} \quad . \tag{23}$$

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

$$\widehat{i} = \underset{i}{\operatorname{argmax}}\, d_i \quad . \tag{24}$$

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

$$\begin{array}{ll} \text{first object} & (\kappa_1, \widehat{\boldsymbol{\phi}}_{\kappa_1}, \widehat{\boldsymbol{t}}_{\kappa_1}) \\ \text{second object} & (\kappa_2, \widehat{\boldsymbol{\phi}}_{\kappa_2}, \widehat{\boldsymbol{t}}_{\kappa_2}) \\ \vdots & \\ \text{last object} & (\kappa_{\widehat{i}}, \widehat{\boldsymbol{\phi}}_{\kappa_{\widehat{i}}}, \widehat{\boldsymbol{t}}_{\kappa_{\widehat{i}}}) \end{array} \quad . \tag{25}$$

In order to evaluate the recognition algorithm for multi-object scenes, not only the object classification result Ω_{κ_i} and the object localisation result $(\widehat{\boldsymbol{\phi}}_{\kappa_i}, \widehat{\boldsymbol{t}}_{\kappa_i})$ have to be verified, but also the number $\widehat{i}$ of objects found in the scene $\boldsymbol{f}$ must be checked.

IV. Experiments and Results

This section discusses the performance of our system on 3D object recognition in a real world environment. The image database (3D-REAL-ENV) used in this experiment is described in Section IV-A. Classification and localisation rates for single-object scenes are presented in Section IV-B, while Section IV-C evaluates the system performance for multi-object scenes.

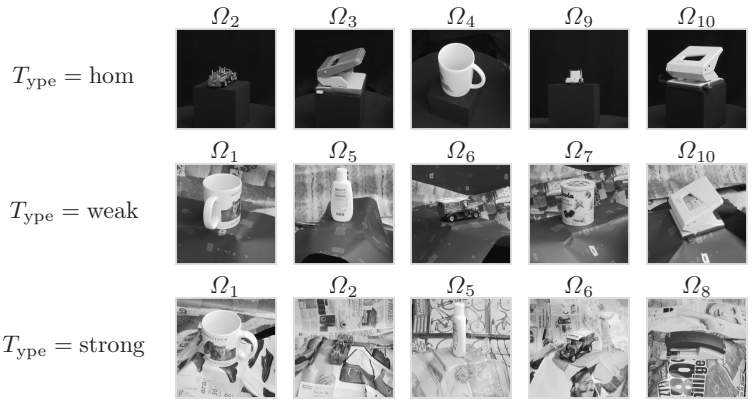

Figure 3. Examples of test scenes on all three types of background $T_{\mathrm{ype}} \in \{\mathrm{hom}, \mathrm{weak}, \mathrm{strong}\}$. The top row shows images with homogeneous background ($T_{\mathrm{ype}} = \mathrm{hom}$), the middle row images with weak heterogeneous background ($T_{\mathrm{ype}} = \mathrm{weak}$), and the bottom row images with strong heterogeneous background ($T_{\mathrm{ype}} = \mathrm{strong}$).

A. 3D-REAL-ENV Image Database

In our experiments we used the 3D-REAL-ENV [30] database consisting of ten real world objects which can be seen in Figure 3. The object pose in 3D-REAL-ENV is defined by internal translations $\boldsymbol{t}_{\mathrm{int}} = (t_x, t_y)^{\mathrm{T}}$ and external rotation parameters $\boldsymbol{\phi}_{\mathrm{ext}} = (\phi_x, \phi_y)^{\mathrm{T}}$. The objects were captured in RGB at a resolution of 640×480 pixels under three different illumination settings $I_{\mathrm{lum}} \in \{\mathrm{bright}, \mathrm{average}, \mathrm{dark}\}$. For this experiment the images were resized to 256×256 pixels.

Training images were captured with the objects against a dark background from 1680 different viewpoints under two different illumination settings $I_{\mathrm{lum}} \in \{\mathrm{bright}, \mathrm{dark}\}$. This produced 3360 training images in total for each 3D-REAL-ENV object. Each object was placed on a turntable performing a full rotation ($0^\circ \leq \phi_{\mathrm{table}} < 360^\circ$) while the camera attached on a robotic arm was moved on a vertical to horizontal arc ($0^\circ \leq \phi_{\mathrm{arm}} \leq 90^\circ$). 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 amounts 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 $\boldsymbol{t}_{\mathrm{int}} = (t_x, t_y)^{\mathrm{T}}$. These translations parameters were determined manually after acquisition.

For testing, the ten 3D-REAL-ENV objects were captured from 288 different viewpoints under the average illumination setting ($I_{\mathrm{lum}} = \mathrm{average}$) and against three different backgrounds: homogeneous, weak heterogeneous, and strong heterogeneous. This resulted into three test sets of 2880 images each denoted according to the background used as $T_{\mathrm{ype}} \in \{\mathrm{hom}, \mathrm{weak}, \mathrm{strong}\}$. Test scenes of the first type ($T_{\mathrm{ype}} = \mathrm{hom}$) were taken on homogeneous black background, while 200 different real backgrounds were used to create heterogeneous backgrounds ($T_{\mathrm{ype}} \in \{\mathrm{weak}, \mathrm{strong}\}$). In scenes with weak heterogeneous background ($T_{\mathrm{ype}} = \mathrm{weak}$) the objects are easier to distinguish from the background than in scenes where strong heterogeneous background ($T_{\mathrm{ype}} = \mathrm{strong}$) have been used (see Figure 3). Similarly to the acquisition of training images, the objects were put on a turntable ($0^\circ \leq \phi_{\mathrm{table}} < 360^\circ$) and the camera moved on a robotic arm from vertical to horizontal ($0^\circ \leq \phi_{\mathrm{arm}} \leq 90^\circ$). However, for test images the turntable's rotation between two successive steps is 11.25°. Thus, test views are different from the views used for training. Also, the illumination in the test scenes is different from the illumination in the training images.

3D-REAL-ENV		Classification Rate [%]			Localisation Rate [%]		
		Hom. Back.	Weak Het.	Strong Het.	Hom. Back.	Weak Het.	Strong Het.
4.5°	GL	100	92.2	54.1	99.1	80.9	69.0
	C	100	88.0	82.3	98.5	77.8	73.6
9.0°	GL	100	92.4	55.4	98.7	80.0	67.2
	C	100	88.3	81.2	98.2	76.4	72.1
13.5°	GL	99.4	89.7	56.2	96.9	78.6	65.4
	C	99.6	82.7	80.3	94.9	68.4	66.6
18.0°	GL	99.9	89.2	55.1	96.6	71.4	54.5
	C	97.3	80.6	68.6	94.3	64.9	60.7
22.5°	GL	99.4	86.0	52.8	94.5	60.7	38.6
	C	94.7	74.8	59.2	89.4	52.2	46.2
27.0°	GL	96.5	69.4	54.4	83.8	49.9	32.8
	C	93.8	53.6	50.2	78.3	35.8	35.6

Table I
CLASSIFICATION AND LOCALISATION RATES OBTAINED FOR 3D-REAL-ENV IMAGE DATABASE WITH GRAY LEVEL (GL) AND COLOUR (C) MODELLING. THE DISTANCE OF TRAINING VIEWS VARIES FROM 4.5° TO 27° IN 5 STEPS. FOR EXPERIMENTS, 2880 TEST IMAGES WITH HOMOGENEOUS, 2880 TEST IMAGES WITH WEAK HETEROGENEOUS, AND 2880 IMAGES WITH STRONG HETEROGENEOUS BACKGROUND WERE USED.

B. Experimental Results for Single-Object Scenes

The recognition algorithm for single-object scenes described in Section III-A was evaluated for the 3D-REAL-ENV image database presented in the previous section. The training of statistical object models was performed for 6 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 I. 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 and weak heterogeneous background the recognition algorithm performs well even for gray

	Without Context Modelling			With Context Modelling		
	Hom	Weak	Strong	Hom	Weak	Strong
ON	100%	83.9%	43.2%	99.9%	88.2%	59.2%
CL	100%	91.9%	62.9%	100%	97.0%	87.5%
LL	99.7%	81.7%	58.1%	99.7%	81.7%	58.1%

Table II
QUANTITATIVE COMPARISON OF THE SYSTEM'S PERFORMANCE WITH AND WITHOUT CONTEXT MODELLING. ON - OBJECT NUMBER DETERMINATION, CL - CLASSIFICATION, LL - LOCALISATION.

level modelling. For these types of background the use of computational demanding colour information can be avoided. Object recognition takes 3.6s in one gray level image and 7s in one colour image on a workstation equipped with a Pentium 4, at 2.66 GHz, and 512 MB of RAM.

C. 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 (Υ_1 = kitchen, Υ_2 = nursery, Υ_3 = 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 were trained as described in Section II-F.

Altogether 3240 gray 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 Figure 3. Similarly to the case of single-object scenes, the test images were divided into three types: 1080 images with homogeneous background, 1080 scenes with weak heterogeneous background, and 1080 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 II. 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 increase significantly when using context modelling for scenes with real heterogeneous background.

D. 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 COIL image database (Columbia Object Image Library). COIL-20 presented in [31] consists of 20 objects, while COIL-100 [32] is a completion of COIL-20 with additional 80 objects. Although the COIL image database provides only gray 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. In [33], five tree-based machine learning methods for object classification based on random extraction and classification of subwindows are compared to each other using the COIL-100 dataset. The average classification rate for these approaches amounts to 86.7%.

V. REAL WORLD APPLICATION SCENARIOS

A. Annotation of Museum Visit Photos

It often happens that after spending few hours in a museum we only remember some of the most impressive artefacts on display. 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.

Research initiatives such as SCULPTEUR [3] [34] and CHIP [4] [35] have targeted innovative ways of bringing the benefits of digital technology for preservation, study and protection of heritage collections. Recently, radio frequency identification (RFID) tags have also been used to guide visitors through discovery tours in museums and to provide enhanced information on the items of interest to visitors [36]. Although less interactive than solutions using radio tags the image-based recognition of artefacts is less expensive considering that RF tagged collections need to provide visitors with wireless PDA devices that trigger these tags. Furthermore, it is not encumbered by privacy concerns since the interests of visitors cannot be traced without their consent as in the case of RF tags.

Our targeted application starts from the observation that many museum visitors actually take photos of items on display. As the time goes by they remember less and less information about the artefacts in the photos. In order to enrich the visit experience the museum can provide the visitors with an on-line or on-site service in which a visitor presents a set of digital photos taken inside the museum and the museum returns additional information about the artefacts contained in the photos. Due to the amount of photos that would be presented for annotation such an application is feasible for museums only when the annotation process is entirely automated.

The crucial bottleneck in the automatic annotation system corresponds to 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 quite diverse, being taken at various positions around the artefact display. The scales at which the artefacts appear in various photos also vary according to the distance to the camera and the zoom level used when the photo was captured. However the lighting conditions are mostly invariant and known deriving from the light provided in the museum exhibit space. 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. The lighting

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

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

could be constrained to be similar to that in the room where the artefact is exhibited. Each photo to be recognised is then matched to multiple-views of artefacts in the collection captured in controlled conditions. A multi-scale approach can deal with scale variations. We are currently designing and building an prototype end-user application for this scenarios 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%.

B. 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) [37]. The aim of this analysis is monitoring the quality process in the steel plant. 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.

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 that, 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

$$p_\kappa = \prod_{m=1}^{m=M} p(\boldsymbol{c}_m|\boldsymbol{\mu}_{\kappa,m}, \boldsymbol{\sigma}_{\kappa,m}) \quad , \tag{26}$$

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 (26) is stored in the statistical concept model $\mathcal{M}_\kappa$. These density values are then maximised with Maximum Likelihood (ML) Estimation [38]

$$\widehat{\kappa} = \underset{\kappa}{\operatorname{argmax}} p_\kappa \quad . \tag{27}$$

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

We tested our approach on 240 example metallography images categorised into four quality classes by a human expert. Our system provided the same classification results in 223 cases which yields a classification rate of 92.9%. 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.

VI. Conclusions

This article presents a system for 3D texture-based probabilistic object classification and localisation and its applications. In contrast to shape-based approaches, texture-based methods do not use any segmentation techniques for feature extraction. The features are computed directly from the image pixels as described in Section I.

The training mode of the system (Section II) starts with the local feature extraction by the discrete wavelet transform. Subsequently, a tightly enclosing object area is learned for each object class. Feature vectors inside this object area are represented by normal density functions, while background features are modelled with the uniform distribution. Finally, context dependencies between objects are modelled in the training phase.

The recognition mode of the system is described in Section III. At first we present an approach that deals with single-object scenes and solves the recognition problem by the maximum likelihood estimation. The second recognition algorithm addressed in this paper deals with the problem of object classification and localisation in multi-object scenes. However, it takes into consideration context dependencies between objects, which are statistically modelled in the training phase.

As can be seen in (18), in order to perform the classification and localisation of a single object in a single image the density values $p_{\kappa,h}$ are compared to each other for all objects Ω_κ and for all pose hypotheses $(\boldsymbol{\phi}_h, \boldsymbol{t}_h)$. However, the number of objects N_Ω and the number of pose hypotheses N_h might vary depending on the task definition and the desired localisation accuracy. The running time of the recognition algorithm T_{rec} highly depends on these numbers and can be expressed by $T_{\mathrm{rec}} \sim N_\Omega \cdot N_h$.

Experimental investigation has been carried out (Section IV) on an image database of over 40000 images specifically recorded for 3D object recognition in a real world environment (3D-REAL-ENV). The classification and localisation results obtained in the experiment prove the high performance of our system. A boost in performance is obtained by using colour and context modelling. The classification rate achieved for 3D-REAL-ENV test images with strong heterogeneous background is 54.1% for gray level modelling while when colour information is applied the classification reaches 82.3%. The performance of the localisation algorithm is also improved by colour modelling for difficult heterogeneous environments from 69.0% on gray level modelling to 73.6% on colour modelling. Furthermore, due to the modelling of context dependencies between objects, higher classification rates where obtained for multi-object scenes. The classification rate for multi-object scenes with strong heterogeneous background but without considering context dependencies amounts to 62.9%, while taking into account context increases the classification rate to 87.5%.

The system described in this paper is currently being embedded in real applications (Section V). The first application

targeted is recognition of museum artefacts from photos taken by visitors. The second application investigated is the analysis of metallography images from an steel plant.

As shown, the texture-based statistical object classification approach presented in this article can be easily adapted to other computer vision tasks. Two such tasks, namely classification of museum artefacts and of metallography images are described in here. Improvements are possible and we are currently investigating some promising paths. One extension of our approach is 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 intend to apply it for image and video content retrieval.

ACKNOWLEDGEMENTS

Research activities leading to this work have been supported by the European Commission under the contract FP6-027026-K-SPACE.

REFERENCES

[1] L. J. Latecki, R. Lakaemper, and D. Wolter, "Optimal partial shape similarity," *Image and Vision Computing Journal*, vol. 23, pp. 227–236, 2005.

[2] B. Schiele and J. L. Crowley, "Recognition without correspondence using multidimensional receptive field histograms," *International Journal of Computer Vision*, vol. 36, no. 1, pp. 31–50, January 2000.

[3] I. Ulusoy and C. M. Bishop, "Generative versus discriminative methods for object recognition," in *International Conference on Computer Visions and Pattern Recognition (Volume 2)*. San Diego, USA: IEEE Computer Society, June 2005, pp. 258–264.

[4] I. T. Jolliffe, *Principal Component Analysis*. Springer, 2002.

[5] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, 2001.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[7] P. M. Roth and M. Winter, "Survey of appearance-based methods for object recognition," Inst. for Computer Graphics and Vision, Graz University of Technology, Austria, Tech. Rep. ICG-TR-01/08, 2008.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, 2000.

[9] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[10] M. Pontil and A. Verri, "Support vector machines for 3d object recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 6, pp. 637–646, January 1998.

[11] Y. Freund and R. E. Shapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer System Sciences*, vol. 55, pp. 119–139, 1997.

[12] R. Gross, I. Matthews, and S. Baker, "Appearance-based face recognition and light-fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 4, pp. 449–465, April 2004.

[13] H. Schneiderman and T. Kanade, "Object detection using the statistics of parts," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 151–177, March 2004.

[14] C. H. Park and H. Park, "Fingerprint classification using fast fourier transform and nonlinear discriminant analysis," *Pattern Recognition*, vol. 38, no. 4, pp. 495–503, April 2005.

[15] L. Heutte, A. Nosary, and T. Paquet, "A multiple agent architecture for handwritten text recognition," *Pattern Recognition*, vol. 37, no. 4, pp. 665–674, April 2004.

[16] M. Zobel, J. Denzler, B. Heigl, E. Nöth, D. Paulus, J. Schmidt, and G. Stemmer, "Mobsy: Integration of vision and dialogue in service robots," *Machine Vision and Applications*, vol. 14, no. 1, pp. 26–34, April 2003.

[17] C. H. Li and P. C. Yuen, "Tongue image matching using color content," *Pattern Recognition*, vol. 35, no. 2, pp. 407–419, February 2002.

[18] H. Y. Ngan, G. K. Pang, S. Yung, and M. K. Ng, "Wavelet based methods on patterned fabric defect detection," *Pattern Recognition*, vol. 38, no. 4, pp. 559–576, April 2005.

[19] J. Gausemeier, M. Grafe, C. Matysczok, R. Radkowski, J. Krebs, and H. Oelschlaeger, "Eine mobile augmented reality versuchsplattform zur untersuchung und evaluation von fahrzeugergonomien," in *Simulation und Visualisierung*, T. Schulze, G. Horton, B. Preim, and S. Schlechtweg, Eds. Magdeburg, Germany: SCS Publishing House e.V., March 2005, pp. 185–194.

[20] Y. Amit, D. Geman, and X. Fan, "A coarse-to-fine strategy for multi-class shape detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 12, pp. 1606–1621, December 2004.

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

[22] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing visual features for multiclass and multiview object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 854–869, May 2007.

[23] Y. Jin and S. Geman, "Context and hierarchy in a probabilistic image model," in *IEEE Conference on Computer Vision and Pattern Recognition*, New York, USA, June 2006, pp. 2145–2152.

[24] B. Leibe and B. Schiele, "Analyzing contour and appearance based methods for object categorization," in *IEEE Conference on Computer Vision and Pattern Recognition*, Madison, USA, June 2003.

[25] D. G. Lowe, "Object recognition from local scale-invariant fearures," in *7. International Conference on Computer Vision (ICCV)*, Corfu, Greece, September 1999, pp. 1150–1157.

[26] S. Mahamud and M. Hebert, "The optimal distance measure for object detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Madison, USA, June 2003.

[27] B. Heigl, *Plenoptic Scene Modeling from Uncalibrated Image Sequences*. Stuttgart, Germany: ibidem-Verlag, 2004.

[28] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, July 1989.

[29] M. Grzegorzek, M. Reinhold, and H. Niemann, "Feature extraction with wavelet transformation for statistical object recognition," in *4th International Conference on Computer Recognition Systems*, M. Kurzynski, E. Puchala, M. Wozniak, and A. Zolnierek, Eds. Rydzyna, Poland: Springer-Verlag, Berlin, Heidelberg, May 2005, pp. 161–168.

[30] M. Grzegorzek and H. Niemann, "Statistical object recognition including color modeling," in *2nd International Conference on Image Analysis and Recognition*, M. Kamel and A. Campilho, Eds. Toronto, Canada: Springer-Verlag, Berlin, Heidelberg, LNCS 3656, September 2005, pp. 481–489.

[31] S. Nene, S. Nayar, and H. Murase, "Columbia object image library (coil-20)," Department for Computer Science, Columbia University, Tech. Rep. Technical Report CUCS–005–96, 1996.

[32] S. Nene, S. Nayar, and Murase, "Columbia object image library (coil-100)," Department for Computer Science, Columbia University, Tech. Rep. Technical Report CUCS–006–96, 1996.

[33] R. Maree, P. Geurts, J. Piater, and L. Wehenkel, "Decision trees and random subwindows for object recognition," in *ICML Workshop on Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, August 2005.

[34] S. Goodall, P. Lewis, K. Matrinez, P. Sinclair, F. Giorgini, M. Addis, M. Boniface, C. Lahanier, and J. Stevenson, "Sculpteur: Multimedia retrieval for museums," in *Third International Conference on Image and Video Retrieval (CIVR 2004)*, Dublin, Ireland, April 2004, pp. 638–646.

[35] L. Aroyo, Y. Wang, R. Brussee, P. Gorgels, L. Rutledge, and N. Stash, "Personalized museum experience: The rijksmuseum use case," in *In Proceedings of Museums and the Web*, San Francisco, USA, April 2007.

[36] T. Liu, T. Tan, and Y. Chu, "The ubiquitous museum learning environment: Concept, design, implementation, and a case study," in *Sixth International Conference on Advanced Learning Technologies*, Kerkrade, The Netherlands, July 2006, pp. 989–991.

[37] P. Praks, M. Grzegorzek, R. Moravec, L. Valek, and E. Izquierdo, "Wavelet and eigen-space feature extraction for classification of metallography images," in *European-Japanese Conference on Information Modeling and Knowledge Bases*, H. Jaakkola, Y. Kiyoki, and T. Tokuda, Eds. Pori, Finland: Juvenes Print-TTY, Tampere, June 2007, pp. 193–202.

[38] A. R. Webb, *Statistical Pattern Recognition*. Chichester, UK: John Wiley & Sons Ltd, 2002.

AUTHOR BIOGRAPHIES

Marcin Grzegorzek received his PhD with distinction in the field of statistical object recognition from the University of Erlangen-Nuremberg in 2007. Then, he was Research Assistant at the Queen Mary, University of London and worked, in general, on pattern recognition and multimedia analysis. Currently, Marcin is employed at the University of Koblenz-Landau and his scientific investigations concentrate on semantically driven image analysis and cross-media technologies. Marcin is author of 6 journal articles, 15 conference papers, and a text book. Moreover, he is a Guest Editor of the International Journal on Multimedia Tools and Applications.

Sorin Sav obtained his PhD from Dublin City University in 2005 for a thesis based on using video objects and relevance feedback in content-based retrieval. He is currently a postdoctoral researcher in the Centre for Digital Video Processing, a 45 person multi-disciplinary research centre based in Dublin City University, Ireland. His research interests include image/video analysis, content-based retrieval and design of novel applications for interactive TV. Since 2005 he has published 15 peer-reviewed papers in international conferences and worked closely with a variety of multinational industry partners.

Ebroul Izquierdo is a professor (chair) of multimedia and computer vision and head of the Multimedia and Vision Group at Queen Mary, University of London . For his thesis on the numerical approximation of algebraic-differential equations, he received the Dr. Rerun Naturalium (PhD) from the Humboldt University , Berlin, Germany, in 1993. From 1990 to 1992 he was a teaching assistant at the department of applied mathematics, Technical University Berlin. From 1993 to 1997 he was with the Heinrich-Hertz Institute for Communication Technology, Berlin , Germany, as associated researcher. From 1998 to 1999 he was with the Department of Electronic Systems Engineering of the University of Essex as a senior research officer. Since 2000 he has been with the Electronic Engineering department, Queen Mary, University of London.

Prof. Izquierdo is an associate editor of the IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) and the EURASIP journal on image and video processing. He has served as guest editor of three special issues of the IEEE TCSVT, a special issue of the journal Signal Processing: Image Communication and a special issue of the EURASIP Journal on Applied Signal Processing.

Prof. Izquierdo is a Chartered Engineer, a Fellow of the The Institution of Engineering and Technology (IET), chairman of the Executive Group of the IET Visual Engineering Professional Network, a senior member of the IEEE, a member of the British Machine Vision Association and a member of the steering board of the Networked Audiovisual Media technology platform of the European Union. He is member of the programme committee of the IEEE conference on Information Visualization, the international program committee of EURASIP&IEEE conference on Video Processing and Multimedia Communication and the European Workshop on Image Analysis for Multimedia Interactive Services. Prof. Izquierdo has served as session chair and organiser of invited sessions at several conferences.

Prof. Izquierdo coordinated the EU IST project BUSMAN on video annotation and retrieval. He is a main contributor to the IST integrated projects aceMedia and MESH on the convergence of knowledge, semantics and content for user-centred intelligent media services. Prof. Izquierdo coordinates the European project Cost292 and the FP6 network of excellence on semantic inference for automatic annotation and retrieval of multimedia content, K-Space.

Prof. Izquierdo has published over 300 technical papers including chapters in books.

Noel E. O'Connor graduated from Dublin City University (DCU) with a B.Eng. in Electronic Engineering (1992) and a PhD (1998), after working for 2 years as a research assistant for Teltec Ireland. He is currently an Associate Professor in the School of Electronic Engineering and a PI in CLARITY: Centre for Sensor Web Technologies. Since 1999 he has published over 130 peer-reviewed publications, made 11 standards submissions, filed 5 patents and spun off a campus company, Aliope Ltd. He has acted as PC Chair for 3 international conferences and regularly reviews for a number of respected journals and acts as a PC member for many international conferences. His current research interests include scene-level classification, multi-spectral video analysis, smart AV sensored environments, and 2D/3D visual capture. He is a member of the IEEE, Engineers Ireland and the Institution of Engineering and Technology.