DENSE STATISTIC VERSUS SPARSE FEATURE-BASED APPROACH FOR 3D OBJECT RECOGNITION

P. Decker[1], S. Thierfelder[1], D. Paulus[1], and M. Grzegorzek[2,]

[1] Research Group for Active Vision, University Koblenz-Landau
Universitaetsstr. 1, 56070 Koblenz, Germany
{decker,susanne.thierfelder,paulus@uni-koblenz.de}

[2] Research Group for Pattern Recognition, University of Siegen
Hoelderlinstr. 3, 57076 Siegen, Germany
marcin.grzegorzek@uni-siegen.de

In this article we introduce and compare two approaches towards automatic classification of 3D objects in 2D images. The first one is based on statistical modeling of wavelet features. It estimates probability density functions for all possible object classes considered in a particular recognition task. The second one uses sparse local features. For training, SURF features are extracted from the training images. During the recognition phase, features from the image are matched geometrically, providing the best fitting object for the query image. Experiments were performed for different training sets using more than 40.000 images with different backgrounds. Results show very good classification rates for both systems and point out special characteristics for each approach, which make them more suitable for different applications.

Introduction

One of the most fundamental problems of computer vision is the classification of objects in digital images. The task of object classification is to determine the classes of objects occurring in an image from a set of predefined object classes. Generally, the number of objects in a scene is unknown, however, in this work we assume that exactly one object is expected in an image.

In this work we introduce and compare two different approaches for classification of 3D objects in digital images. The first one is based on statistical modeling of local wavelet features and uses maximum likelihood estimation to determine the class of the object in a scene. The second approach is based upon robust local point descriptors. Due to advances in this field during the last decade [1, 2] these descriptors increase to play an important role in object recognition and similar fields of research. Their invariance properties towards change of illumination, scale and orientation make them perfect candidates for robust object recognition systems.

The contribution of the paper lies in (i) the extension of the system for statistical object recognition by including LAB color space modeling for feature extraction, (ii) the introduction of a completely new algorithm for object recognition based upon robust local point descriptors, and (iii) a comprehensive comparative evaluation of these approaches leading to interesting scientific conclusions.

The paper is structured as follows. Section 2 describes shortly the system for statistical object recognition. Section 3 gives an overview over the object classification method using local point descriptors. In Section 4, a comprehensive quantitative comparison of these two approaches performed for more than 40.000 images is given. Finally, Section 5 closes the paper by providing some interesting conclusions and plans for the future work in this research area.

Dense Statistical Object Classification

Our framework performs the supervised statistical learning in following steps: (i) object acquisition from different viewpoints, (ii) preprocessing into one of the investigated color spaces, (iii) feature extraction, (iv) object area definition, and (v) estimation of the multivariate likelihood density function.

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. In this way all objects are presented to the systems from different known viewpoints. The original images are then resized to 256×256 pixels. The system is able to work with gray level, RGB, and Lab images.

Subsequently, the system determines a set of local feature vectors for all preprocessed training images. For this, the images are divided into neighborhoods of size 8×8 pixels and three steps of the wavelet transform are performed for all of them. Each such neighborhood is then represented by the following feature vector

$$\mathbf{c}_m = \begin{pmatrix} c_{m,1} \\ c_{m,2} \end{pmatrix} = \begin{pmatrix} \ln(2^s|b_s|) \\ \ln[2^s(|d_{0,s}| + |d_{1,s}| + |d_{2,s}|)] \end{pmatrix} \quad . \quad (1)$$

The first element is computed from the low-pass coefficient, the second element results from both, the low-pass and the high-pass filtering. The parameter s denotes the number of steps of the wavelet transform and is in our case equal to 3 (neighborhoods of size 8×8 pixels). For RGB and LAB images six-dimensional local feature vectors are computed in the same way, whereas the color channels are treated separately from each other.

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 feature vectors describing the object are considered for statistical object modeling. Since the object usually composes a part of the image, a tightly enclosing bounding region O called object area is defined for each object class. For clarity, we will use the term object area to actually refer to the set of features belonging to the object.

In order to handle illumination changes and low-frequency noise, the elements of the local feature vectors are interpreted as random variables. Assuming the object's feature vectors as statistically independent of the feature vectors outside the object area, the background feature vectors can be disregarded here. The object feature vectors are represented by Gaussian density functions

$$p(\mathbf{c}_m) = p(\mathbf{c}_m|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m) \quad . \quad (2)$$

Assuming the statistical independence of the elements of the feature vectors, as well as of the feature vectors inside the object area, the multivariate likelihood density function describing an object class can be written as follows

$$p(O|\mathbf{B}) = \prod_{\mathbf{c}_m \in O} p(\mathbf{c}_m|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m) \quad . \quad (3)$$

$\boldsymbol{B}$ comprises the mean value vectors and the standard deviation vectors.

In the recognition phase, a set of feature vectors O is determined from a test image and evaluated against density functions of all regarded object classes. As classification result the class with the highest probability value is understood

$$\hat{\kappa} = \underset{\kappa}{\operatorname{argmax}}\, p(O|\mathbf{B}_\kappa) \quad . \quad (4)$$

Sparse Feature-Based Object Classification

This section describes the extraction which is applied to both, training and test images, as well as the hypothesis generation which is equivalent to the recognition step and uses the extracted features.

As feature detector and descriptor we use SURF [2]. SURF is a point feature detector which also provides a descriptor for matching. Its main advantage is the fast computation while the features are distinctive enough to enable robust object recognition even under difficult circumstances such as partial occlusion and cluttered background. SURF uses a scale space approach [3]. The scale extends the image domain by a third

dimension σ, where σ is the standard deviation of a Gaussian smoothing kernel applied to the image. A key-point k in scale space is detected if the determinant of its Hessian is a maximum with respect to its 26 neighbors in the three dimensional scale space. The scale space approach makes the detection step scale invariant, e. g. key-points on an object are detected in the same locations, regardless the scaling of the object in the image.

Each key-point is assigned an orientation θ to enable rotation invariance. A Haar wavelet is used to approximate gradient information in a circular region around the key-point. The responses are weighted with a Gaussian centered at the key-point and represented as 2D vectors, which are ordered by angle. Then a sliding window of size $\pi/3$ is used to detect a maximum which is then used as orientation θ for the key-point. To make key-points recognizable, a descriptor δ is assigned to each key-point. Therefore, a square region aligned with the key-point orientation θ is considered. This region is divided in 4×4 subregions. Each of the subregions provides four values, which are the responses to a Haar wavelet function in x and y direction and their absolute values, sampled at 25 positions within the subregion. The result is a descriptor vector δ of length 64, describing intensity changes in the neighborhood of the key-point. To sum up, we extract from each image a number of SURF features f. A feature is a tuple $f = (x, y, \sigma, \theta, \delta)$ containing the position, scale and orientation of the feature in the image, as well as a descriptor. The features are invariant towards scaling, position in the image and in-plane rotations. They are also robust towards changes in illumination and lesser off-plane rotations.

Features between a training image and a test image are matched by their descriptors. To find the most similar descriptor vector to a given one, a distance measure is needed. Since simple distance thresholds do not perform good in high dimensional space, Lowe introduced the distance ratio [1]. The distance ratio is the ratio of the distance to the best fitting and the second best fitting descriptor. If this quotient is low enough, the best fit is considered a matching feature. If the quotient is higher than a given threshold, it means that the best and the second best descriptor fit almost equally well. This leads to the assumption, that they are very likely the best matches just by chance, and not because one of them actually matches the query descriptor. The distance ratio also sorts out ambiguous matches, which may result from repeating textures on objects.

For the fast nearest neighbor and distance computation in the high dimensional descriptor space, we use an approximate nearest neighbor approach [4]. The result of feature matching is a set of matches between features extracted from a training and the test image. This set may still contain outliers, i. e. matches between features which do not correspond to the same world point.

Each feature match gives a hint of how the object could be visible in the test image. The position, scale and orientation of a feature in the test image in relation to it's known position, scale and orientation on the object define a position, scale and orientation of the object in the image. To cluster this information from all feature matches, a four dimensional Hough space over possible object positions (x, y, σ, θ) is created. The bins size for the Hough space can be chosen quite liberally, also to avoid getting too many bins. We chose 10 bins for each dimension, resulting in 104 bins describing different possible object positions in the image. Each match then votes into 16 bins (the one it falls into and the closest ones of each dimension to avoid discretization effects). For the next step, only bins with at least five entries are considered.

For each bin passing the Hough Clustering step, RANSAC [5] is employed to identify the best homography for the set of correspondences. To avoid processing too many bins, we order them by their number of entries. We terminate when the next bin could not possibly get a greater number of inliers than the best homography so far. The result is a homography describing the relative position, orientation and scale of the best fitting training image for a test image, as well as the number of features supporting this hypothesis.

Since there is no rejection class, each test image has to be assigned exactly one object class it belongs to. In our case, we chose the class from which a training image scored a homography with the overall top number of inliers.

EXPERIMENTS AND RESULTS

For our experiments we used the 3D-REAL-ENV image database consisting of ten objects taken on real heterogeneous background. For more details please refer to [6].

Both methods presented in this paper have been evaluated for 6 different sets of training images. The training sets differ in the distance of adjacent training views (4.5◦, 9◦, 13.5◦, 18◦ , 22.5◦, 27◦). Moreover, experiments have been performed for the three kinds of test images, test images with homogeneous, weak heterogeneous, and strong heterogeneous background. The classification rates for both approaches are summarized in the following table

Distance of Training Views [°]	Approach	Classification Rate [%]		
		Hom. Back.	Weak Het.	Strong Het.
4.5	Sparse	95.6	91.3	74.3
	Dense	100	88.0	82.3
9.0	Sparse	95.9	92.0	76.4
	Dense	100	88.3	81.2
13.5	Sparse	94.9	90.8	73.9
	Dense	99.6	82.7	80.3
18.0	Sparse	93.9	90.6	73.4
	Dense	97.3	80.6	68.6
22.5	Sparse	92.0	87.0	69.3
	Dense	94.7	74.8	59.2
27.0	Sparse	92.3	87.6	65.6
	Dense	93.8	53.6	50.2

CONCLUSIONS

In this paper we introduce and experimentally compare two different approaches for classification of 3D objects in digital images. The first one (see Section 2) is based on statistical modeling of local wavelet features and uses the maximum likelihood estimation to determine the class of the object in a scene. The second approach (see Section 3) is based upon robust local point descriptors. It uses generalized Hough Transform Clustering and a RANSAC framework for homography estimation between training and test images. The contribution of the paper lies in (i) the extension of the system for statistical object recognition by including LAB color space modeling for feature extraction, (ii) the introduction of a completely new algorithm for object recognition based upon robust local point descriptors, and (iii) a comprehensive comparative evaluation of these approaches leading to interesting scientific conclusions.

The results show that the dense statistical approach for object classification (Section 2) brings better results in very difficult environments (strong heterogeneous background) when the number of training images is high. However, it is highly dependent on the number of training images, while the classification rate for the sparse feature-based approach remains almost constant for varying training sets.

In the future, we will extend the systems towards object localization and comprehensively evaluate their performance for the object pose estimation task.

References

1. LOWE D. G.: Distinctive image features from scale-invariant key-points. International Journal of Computer Vision, 60, 2 (2004), 91–110.
2. BAY H., TUYTELAARS T., VAN GOOL L.: Surf: Speeded up robust features. ECCV (2006), 404–417.
3. LINDEBERG T.: Scale-Space Theory in Computer Vision. Kluwer Academic Publishers, Dordrecht, Netherlands, 1994.
4. MUJA M.: Flann, fast library for approximate nearest neighbors, 2009. http://mloss.org/software/view/143/.
5. FISCHLER M. A., BOLLES R. C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24, 6 (1981), 381–395.
6. GRZEGORZEK M., NIEMANN H.: Statistical object recognition including color modeling. In 2nd International Conference on Image Analysis and Recognition (Toronto, Canada, September 2005), Kamel M., Campilho A., (Eds.), Springer-Verlag, Berlin, Heidelberg, LNCS 3656, pp. 481–489.