A *Bag of Features* Approach for 3D Shape Retrieval

Janis Fehr[1,2], Alexander Streicher[2] and Hans Burkhardt[2]

[1] HCI, University Heidelberg, Germany
[3] LMB, University Freiburg, Germany
janis.fehr@iwr.uni-heidelberg.de

Abstract. In this paper, we present an adaptation the *Bag of Features* (BoF) concept to 3D shape retrieval problems. The BoF approach has recently become one of the most popular methods in 2D image retrieval. We extent this approach from 2D images to 3D shapes. Following the BoF outline, we address the necessary modifications for the 3D extension and present novel solutions for the parameterization of 3D patches, a 3D rotation invariant similarity measure for these patches and a method for the codebook generation. We experimentally evaluate the performance of our methods on the *Princeton Shape Benchmark*.

1 Introduction

The retrieval of digitized 3D objects is a rising topic. Similar to 2D image retrieval, which recently has become a very popular research topic, the constantly growing size of available 3D data triggers the need for effective search methods. There have been several practically important applications to 3D object retrieval, such as retrieving 3D protein structures from very large databases in bio-informatics and computational chemistry [1] or the retrieval of 3D objects from depth images (laser range scans) in robot navigation [2].
We apply our methods to a more academic problem setting given by the Princeton Shape Benchmark (PSB) [3], which has become a standard benchmark for 3D shape retrieval.

1.1 Related Work

We limit our brief review of the related work to methods which have been applied to the Princeton Shape Benchmark and thus can be compared to our results later on.
The **Spherical Extent Function (EXT)** [4] projects the distance of the object center to each point of the object surface onto the enclosing outer sphere. The resulting spherical distance map is then expanded in Spherical Harmonics from which the $\mathcal{SH}_{abs}$ (4) feature is extracted. The **Spherical Harmonic Descriptor (SHD)** [5] is very similar to EXT, it also computes $\mathcal{SH}_{abs}$ over several

radii, but organizes the results in a 2D histogram. The **Light Field Descriptor (LFD)** [6] uses multiple 2D views of 3D shapes. Rotation invariance is achieved by a collection of 100 2D views per object, which are rendered orthogonal to the outer enclosing sphere of the object. Then a set of 2D features (mostly geometric and Zernike moments) is computed for each 2D view. Currently, LFD is the best performing approach on the PSB.

All of these methods have in common that they try to model object shapes at a global level which has the disadvantage that the assumption that objects of the same class are sharing the same base shape is not always adequate - especially when one considers more semantic groupings with high intra-class variance as presented by the PSB. In 2D image retrieval, these problems have been approached quite successfully by BoF methods (see section 2). Hence, there have been several previous attempts to introduce a BoF approach for 3D shape retrieval, like [7], using Spin Images as local 3D patch descriptors. However, the results of these attempts were rather poor (see experiments), which we suspect to be mostly due to the limited discrimination power of spin images. We try to overcome these problems by the use of other local 3D patch descriptors.

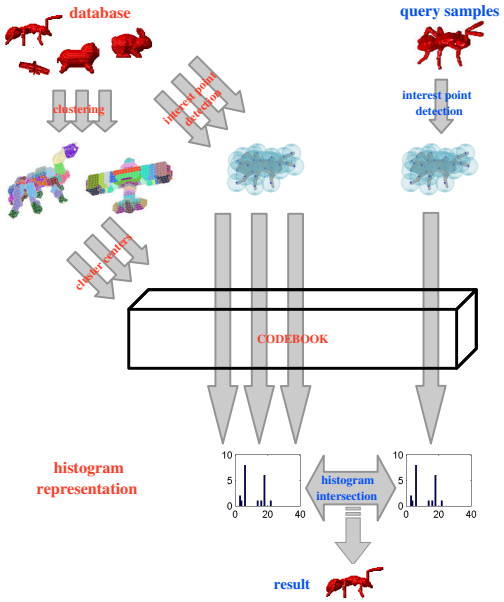


Fig. 1. Schematic overview of the "Bag of Features" concept.

2 3D Shape Retrieval with Local Patches

2.1 The *Bag of Features* Concept

One "state of the art" approach in modern 2D image retrieval is commonly known under the name *Bag of Features* (BoF) [8][9]. The method of BoF is largely inspired by the *Bag of Words* [10] concept which has been used in text retrieval for quite some time.
Even though there are countless variations of retrieval algorithms emerging under the label *Bag of Features* [8][9] and it is hard to capture *the* actual BoF algorithm, there is a common concept which is shared by all of theses methods.

Local Features: The central aspect of the BoF concept is to move away from a global image description and to represent images as a collection of local properties. These local properties are derived in form of (invariant) **image features**, e.g. the very popular SIFT features [11], which are computed on small subimages called **patches**. The patches, are simply small rectangular regions which are extracted around **interest points** (see section 4).
Codebook Representation: The second main aspect of the "Bag of Features" concept is the way images are represented as collections of local features and how two or more of these representations can be compared. The basic idea here is to use a class independent **clustering** over the feature representations of all patches (from all database samples). The representatives of the resulting clusters are then used as entries of a unified (class independent) **codebook** [12]. Each patch is then mapped against the entries of this codebook, such that an image is represented as the **histogram** over the best codebook matches of its patches. The similarity of images can then be obtained by comparing the BoF histograms, e.g. by histogram intersection.
Figure 1 gives a schematic overview of the computation steps in the codebook generation and retrieval stage of the 'Bag of Features" concept.

3 Mathematical Foundations

A key aspect of our 3D extension of the BoF concept is the idea to parameterize the 3D patches as spheres and to provide a fast rotation invariant similarity measure for these spherical patches. The spherical representation is a natural choice for local 3D patches which allows us to rely on the well established mathematical foundations of the *angular momentum theory* [13] to perform all necessary computation in the harmonic domain.

Spherical Harmonics ($\mathcal{SH}$) [13] form an orthonormal base on the 2-sphere. Analogical to the Fourier Transform, any given real valued and continuous signal f on a sphere with its parameterization over the angles Θ, Φ (latitude and longitude of the sphere) can be represented by an expansion in its harmonic coefficients:

$$f(\Theta, \Phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{m=l} \hat{f}_m^l \overline{Y_m^l}(\Theta, \Phi), \tag{1}$$

where l denotes the band of expansion, m the order for the l-th band, $\hat{f}_m^l$ the harmonic coefficients and $\overline{Y_m^l}$ the complex conjugate of the harmonic base functions Y_m^l that are computed as:

$$Y_m^l(\Theta, \Phi) = \sqrt{\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!}} P_m^l(\cos\Theta) \mathrm{e}^{im\Phi}, \tag{2}$$

where P_m^l is the associated Legendre polynomial.

Rotations in $\mathcal{SH}$: Rotations $\mathcal{R}(\varphi,\theta,\psi)f$ in the Euclidean space find their equivalent representation in the harmonic domain in terms of the so called Wigner D-Matrices [13], which form an irreducible representation of the rotation group $\mathcal{SO}(3)$. For each band l, $D^l(\varphi,\theta,\psi)$ (or abbreviated $D^l(\mathcal{R})$) defines a band-wise rotation in the $\mathcal{SH}$ coefficients, using the Euler notation in zyz-convention with $\varphi,\psi \in [0,2\pi[$ and $\theta \in [0,\pi[$ to parameterize the rotations $\mathcal{R} \in \mathcal{SO}(3)$. Hence, a rotation in the Euclidean space can be estimated in the harmonic domain (with a maximum expansion band b_{max}), by

$$\mathcal{R}f \approx \sum_{l=0}^{b_{max}} \sum_{m=-l}^{l} \sum_{n=-l}^{l} D^l_{mn}(\mathcal{R}) \hat{f}^l_m Y^l_m. \tag{3}$$

The $\mathcal{SH}_{abs}$ Feature: A $\mathcal{SH}$ representation of spherical signals raises the demand for a (rotational invariant) similarity measure between two or more signals. A popular choice [5] is to use the band-wise absolute values of the harmonic power-spectrum, which we refer to as $\mathcal{SH}_{abs}$ feature:

$$\mathcal{SH}_{abs}(\hat{f}^l) := \sum_{m=-l}^{m=l} \|\hat{f}^l_m\|. \tag{4}$$

The main drawback of the $\mathcal{SH}_{abs}$ features is that it obtains its rotation invariance by neglecting the phase information. Hence, $\mathcal{SH}_{abs}$ is an incomplete feature which suffers from its ambiguities.

Fast Correlation in $\mathcal{SH}$: We follow a different approach to obtain a rotation invariant similarity measure between harmonic expansions: the full correlation over all possible rotation angles: The full correlation $f\#g : \mathcal{SO}(3) \to \mathbb{R}$ of two signals f and g under the rotation $\mathcal{R} \in \mathcal{SO}(3)$ on a 2-sphere is given as:

$$(f\#g)(\mathcal{R}) := \int_{S^2} f(\mathcal{R}g) \quad d\phi d\theta d\psi. \tag{5}$$

Fehr et. all. [14] proposed a method for a fast computation of (5) in the harmonic domain by use the *Convolution Theorem*. Starting from the substitution of f and g in (5) by their SH expansions

$$(f\#g)(\mathcal{R}) = \sum_{lmn} \overline{D^l_{mn}(\mathcal{R})} \hat{f}^l_m \overline{\hat{g}^l_n}, \tag{6}$$

their method provides the correlation value for each possible ration in a discrete 3D matrix $C^{\#}$ which represents the angular space over the rotation angles (ϕ,θ,ψ):

$$C^{\#} = \mathcal{F}^{-1}(\widehat{C^{\#}}), \tag{7}$$

with

$$\widehat{C^{\#}}(m,h,m') = \sum_{l=0}^{b_{max}} d^l_{mh}(\pi/2) d^l_{hm'}(\pi/2) \widehat{f}^l_m \overline{\widehat{g}^l_{m'}} \tag{8}$$

and $m, h, m' \in \{-l, \ldots, l\}$. The rotation invariant correlation maximum is then simply the maximum value in $C^{\#}$. Please refer to [14] details on (7) and proofs.

Normalized Cross-Correlation: We follow an approach which is widely known from the normalized cross-correlation of 2D images: First, we subtract the mean from both functions prior to the correlation and then divide the results by the variances:

$$(f\#g)_{norm}(\mathcal{R}) := \int_{S^2} \frac{\big(f - \mathbf{E}(f)\big)\big(\mathcal{R}\big(g - \mathbf{E}(g)\big)\big)}{\sigma_f \sigma_g} \quad \sin\Theta d\Phi d\Theta. \tag{9}$$

Analogous to Fourier transform, we obtain the expectation values $\mathbf{E}(f)$ and $\mathbf{E}(g)$ directly from the 0th $\mathcal{SH}$ coefficient. The variances σ_f and σ_g can be estimated from the band-wise energies:

$$\sigma_f \approx \sqrt{\sum_l |\widehat{f}_l|^2}. \tag{10}$$

Discrete Spherical Harmonic Expansions: For practical applications, we need a discrete version of the Spherical Harmonic transform, i.e. we need to obtain the frequency decomposition of 3D signals at discrete positions $\mathbf{x} \in X$ on discrete spherical surfaces $\mathcal{S}$ of radius r:

$$\mathcal{S}[r]\,(\mathbf{x}) := \{\mathbf{x}' \in \mathbb{R}^3 | \quad \|\mathbf{x} - \mathbf{x}'\|_2 = r\}. \tag{11}$$

To obtain the discrete Vectorial Harmonic transformation $\mathcal{SH}\big(\mathcal{S}[r](\mathbf{x})\big)$, we pre-compute discrete approximations $\tilde{Y}_m^l[r]$ of the orthonormal harmonic base functions as:

$$\mathcal{SH}\big(X|_{\mathcal{S}[r](\mathbf{x})}\big)_m^l := \sum_{\mathbf{x_i} \in \mathcal{S}[r](\mathbf{x})} X(\mathbf{x_i}) \tilde{Y}_m^l[r](\mathbf{x_i}). \tag{12}$$

In order to compute the harmonic transformation of the neighborhoods around each voxel of X, we perform a fast convolution of the pre-computed based functions with the discrete input data:

$$\mathcal{SH}[r]\,(X)_m^l = X * \tilde{Y}_m^l[r]. \tag{13}$$

4 Algorithm

Our approach directly follows the *Bag of Features* scheme (see figure 1). With exception of an additional preprocessing, we simply walk through the BoF pipeline step by step and replace 2D specific algorithms with our own 3D methods.

Preprocessing: Prior to the actual BoF pipeline, we apply a series of pre-processing steps to the objects in the PSB database: primarily, we have to render the objects from triangular mesh format to a volume representation where the voxels inside the object are set to 1 and the voxels outside to 0. We use this rendering step to align the object in the geometric center of the volume and to normalize the object size to a fixed height of the object bounding box. Thus, we obtain translation and scale invariant volume representations of the models.

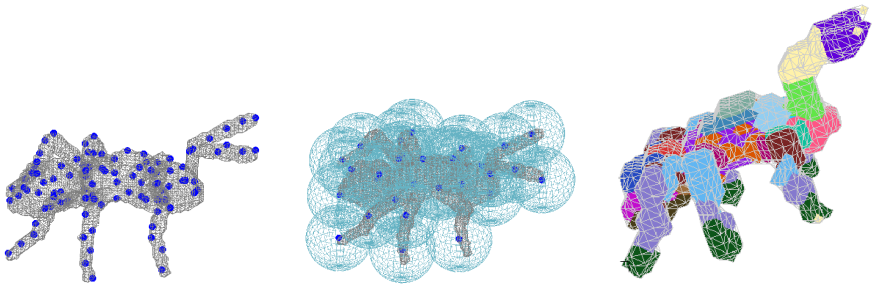
Sampling Points: The next step, and first in the actual BoF pipeline, is to determine the location of the local patches. In the original 2D setting, where the objects of interest are located in more or less cluttered scenes, the detection of interest points is a crucial step: important parts of the target objects should not be missed, while the overall number of interest points directly affects the computational complexity, so that one tries to avoid large numbers of false positive points. For our 3D case, the selection of the interest points is by far less crucial since we have already segmented objects. Hence, we simply apply a simple equidistant sampling on the object surface.

Extracting Local Patches: The next step is to extract the local patches $p(\mathbf{x})$ at the location of the sampling points. In contrast to the original 2D case, were the patches are rectangular areas, we extract spherical patches which are centered in the respective sampling points.
Given the volume rendering of a model X and sampling points at positions $\mathbf{x}$, the associated patches are then represented by a series of m concentric spherical neighborhoods $X|_{\mathcal{S}[r_i](\mathbf{x})}$ (12) at different radii $r_i \in \{r_1, \ldots, r_n\}$. We then expand the local patches radius by radius in Spherical Harmonics. Hence, we define a patch $p(\mathbf{x})$ as collection of radius-wise harmonic expansions up to some upper band $l = b_{max}$ around $\mathbf{x}$:

$$p(\mathbf{x}) := \left\{ \mathcal{SH}\big(X|_{\mathcal{S}[r_1](\mathbf{x})}\big), \ldots, \mathcal{SH}\big(X|_{\mathcal{S}[r_n](\mathbf{x})}\big) \right\}. \tag{14}$$

Figure 3a and 3b illustrate the patch extraction. The motivation to use spherical instead of rectangular patches is obvious considering that we need to obtain full rotation invariance, which often times can be neglected in the case of 2D image retrieval.

Generating the Codebook: While the preprocessing and patch extraction has to be done for all reference and query objects, we now turn to the off-line procedure which is only performed on the initial database. The off-line stage has two different stages: first, we have to generate a problem specific but class independent codebook of local patches, which is done via clustering, and then, we have to represent the database samples in terms of histograms over the codebook.
After the extraction of the patches, we use a simple radius-wise k-means clustering [15] to obtain k patch clusters for each radius of the patch parameterization independently. The key for the clustering is the choice of the similarity function d: we apply the normalized correlation (10) for the Spherical Harmonic domain



(a) (b) (c)

Fig. 3. (a) Extracting spherical patches at the sampling points. (b) extraction of spherical neighborhoods (patches) around the sampling points. (c) Example clustering results where patches were extracted at all object voxels.

to measure the rotation invariant similarity of two patches:

$$d\big(p(\mathbf{x_i}), p(\mathbf{x_j})\big) := p(\mathbf{x_i}) \# p(\mathbf{x_j}). \tag{15}$$

In order to reduce the computational complexity, we do not apply the clustering on all patches from all database samples. Our experiments showed (see 5), that it is sufficient to use a small random subset of 10% of the database to generate a stable set of clusters.
It should be noted, that the class label of the database samples is completely neglected during the clustering since our goal is to obtain a general, class independent representation of local patches in the later codebook. Figure 3c shows example results of the clustering. The final step towards the generation of the codebook is based on the previous clustering. We simply use the cluster centers as representatives in the codebook. Since we perform the clustering for each radius of the patch parameterization independently, we obtain separate codebooks for each radius.

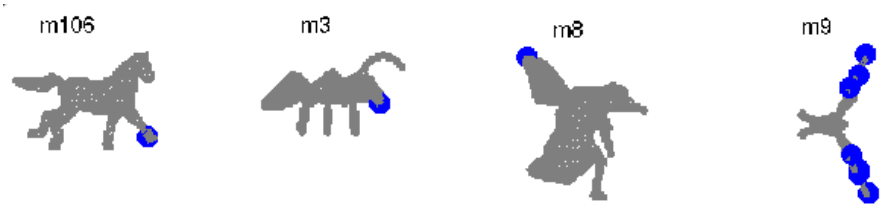


Fig. 4. Sample generalized codebook entry: the figure illustrates the location (blue circle) of a sample codebook entry on several different objects.

4.1 Retrieval by Feature Histograms

After we learned the codebook based on a small subsection of the database, we can pursuit the BoF approach without further changes. As in the original *Bag of Features* concept, all off the database samples are represented as histograms over the codebook entries. We simply use our fast normalized cross-correlation (9) to match all patches of an object with the codebook and rise the count of the histogram bin associated with the best match. Figure 1 illustrates a example codebook histogram representation of an object.

Retrieval Query: Given an query object, we perform the preprocessing and patch extraction steps and then compute its codebook histogram representation just as we do it for the database. We then use a normalized histogram intersection as similarity measure to find the best matches in the database.

5 Experimental Evaluation

We evaluate our proposed approach on the standard PSB experimental setup, as described in [3]. We use the *base* scheme, where the 1814 shapes of the PSB are split into equally large database and query sets.

General Experimental Setup. We use a rendering normalized to the size of 64 voxels on the longest edge of the bounding box (see 4). The codebook is built form a random selection of 10% of the training set (more samples simply increase the computation time without notable effect on the later recognition rate). 8 different radii with $r_i \in \{3, 4, 5, 6, 7, 8, 10, 12\}$ are used to compute the codebook, where b_{max} of the harmonic expansion is increased according with the radius (from 3 to 7). The codebook size k is set to 150 bins per radius.

Given these fixed parameters, we obtained the following results for our approach on the PSB *base* test set: table 5 shows the $k = 1$ nearest neighbor results of our method compared to the results known from literature. Figure 6 shows the precision-recall plots provided by the standardized PSB evaluation. In order to emphasize the use of our $\mathcal{SH}_{corr}$ feature, we additionally implemented a BoF approach where we used the $\mathcal{SH}_{abs}$ feature as patch descriptor.

6 Conclusions and Outlook

The experiments showed that our approach achieves competitive results on the difficult PSB. The main drawback of our method is that we cannot be sure if the given results are actually representing the global optimum of what can be achieved with our method or if we are stuck in a local maximum of the parameter space. Due to the large number of parameters, we face the problem that the maximum search in the parameter space turns out to be quite tedious. A possible further extension of our histogram approach could be to localize the patch positions. Similar to the 2D approach in [16], we could increase the discrimination power by replacing the global histogram with a localized patch histogram.

Features	PSB *base* level
LFD	65.7%†(61.9%)*
BoF with $\mathcal{SH}_{corr}$	**62.4**%
SHD	55.6%†(52.3%)*
EXT	54.9%†
BoF with $\mathcal{SH}_{abs}$	54.5%
BoF with Spin Images	33.5%

Fig. 5. Results for the 3D shape retrieval on the PSB. Results taken from the literature are marked with †, results from our own implementations are marked with ∗. Unfortunately, we were not able to exactly reproduce the results given in the literature. This could be caused by a different initial rendering, which is not discussed in the given literature.

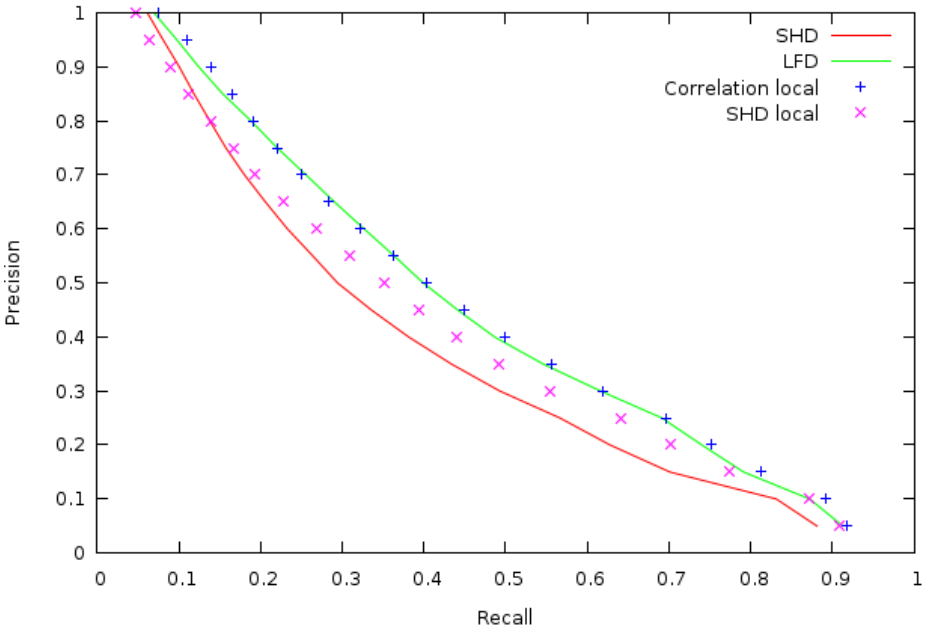


Fig. 6. Precision recall graph for our approach on the PSB *base* test set. The results of our implementation of LFD and SHD reference methods are plotted as lines. We also compare the BoF results with our $\mathcal{SH}_{corr}$ features compared to the use of the $\mathcal{SH}_{abs}$ features as patch descriptor.

References

1. Reisert, M., Burkhardt, H.: Irreducible group representation for 3d shape description. In: Proceedings of the 28th Pattern Recognition Symposium of the German Association for Pattern Recognition (DAGM 2006), Berlin, Germany, LNCS, Springer (2006) 132–142

2. Wolf, J., Burgard, W., Burkhardt, H.: Robust vision-based localization by combining an image retrieval system with Monte Carlo localization. IEEE Transactions on Robotics **21**(2) (2005)
3. Shilane, P., Min, P., Kazhdan, M., Funkhouser, T.: The princeton shape benchmark. In: Shape Modeling International, Genova, Italy. (2004)
4. Saupe, D., Vranic, D.V.: 3d model retrieval with spherical harmonics and moments. In: DAGM 2001. (2001) 392–397
5. Kazhdan, M., Funkhouser, T., Rusinkiewicz, S.: Rotation invariant spherical harmonic representation of 3d shape descriptors. In: Symposium on Geometry Processing. (2003)
6. Chen, D.Y., Ouhyoung, M., Tian, X.P., Shen, Y.T.: On visual similarity based 3d model retrieval. In: Computer Graphics Forum. (2003) 223–232
7. Li, X., Godil, A., Wagan, A.: Spatially enhanced bags of words for 3d shape retrieval. In: Procedings of the ISVC 2008, LNCS 5358. (2008) 349–358
8. Mikolajczyk, K., Leibe, B., Schiele, B.: Local features for object class recognition. In: ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision, IEEE Computer Society (2005) 1792–1799
9. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their location in images. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. Volume 1. (2005) 370–377 Vol. 1
10. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. Mach. Learn. Res. **3** (2003) 993–1022
11. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
12. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. In: ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, IEEE Computer Society (2005) 604–610
13. Brink, D., Satchler, G.: Angular Momentum, Second Edition. Clarendon Press, Oxford (1968)
14. Fehr, J., Reisert, M., Burkhardt, H.: Fast and accurate rotation estimation on the 2-sphere without correspondences. In: Proceedings of the ECCV 2008, LNCS 5303. (2008) 239–253
15. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer (August 2006)
16. Leibe, B., Schiele, B.: Interleaved object categorization and segmentation. In: British Machine Vision Conference (BMVC'03), Norwich, UK, Sept. 9-11. (2003)