Classification of Image Regions Using the Wavelet Standard Deviation Descriptor

Sönke Greve, Marcin Grzegorzek, Carsten Saathoff and Dietrich Paulus
Department of Computer Science, University Koblenz-Landau
Universitätsstraße. 1, 56070 Koblenz, Germany
Emails: {sgreve,marcin,saathoff,paulus}@uni-koblenz.de

***Abstract*—This paper introduces and comprehensively evaluates a new approach for classification of image regions. It is based on the so called *wavelet standard deviation descriptor*. Experiments performed for almost one thousand images with region segmentation given provided reasonable results for a very general application domain: "holiday pictures".**

I. Introduction

In the field of research dealing with image classification and understanding we follow a multi-layered concept. Image data is described in high level and low level semantics. High level semantics are represented by ontologies and low level semantics in feature comparison techniques. This paper is focused on wavelet features following the proposal of the *Wavelet Standard Deviation Descriptor (WSD)* [1] used for the classification of texture patterns. Here it is used in order to classify regions of previously segmented images. The task at hand is to find a label for each region in any given holiday photo that can later be used in image management applications or retrieval systems e.g. to automatically tag inserted photos. The major motivation for this research work lies in the two following statements:

1) finding adequate region labels can improve the research on high-level semantics (narrowing the semantic gap)
2) the problem of image region classification can be reduced to a problem of texture pattern classification

Indications on the first assumption can be found in [2] but in general it is task of the high-level domain. The second assumption is analyzed in detail. Therefore two implementations using the WSD are presented each tested under different modalities. Both classification algorithms compare the WSD of an image region to the WSD of each pre-defined concept. The concepts used are shown in table I. They were trained using a subset of the available images. The region-labels of those images were manually annotated beforehand providing ground truth.

This introduction is followed by a brief overview of similar research in II. Section III introduces the environment of the performed tests followed by the implementation of the mathematical constructs in IV. The results are shown in V with a view on issues, proposals and future developments in VI

II. Related Work

The field of image region classification offers many approaches. In [3] seven feature extraction and comparison methods are compared in their performance of correct image retrieval out of the WANG[1] and the IRMA[2] databases. The evaluated feature extraction methods are image features ($\widehat{=}$ pixel values), color histograms [4], invariant feature histograms [5], Gabor feature histograms [6], Tamura texture feature histograms [7], local features [8] and region based features [9]. As the IRMA database contains too specific medical images, the images of the WANG database get close to the sources of this paper though it is not limited to the loosely defined class of holiday photos. The provided ground truth is also more focused on semantic classes, e.g. "Africa", "food", "monuments", than on elementary texture patterns like "sand", "sky" or "foliage".

In many cases features are generated using an entire image without providing information about specific regions in the image. As soon as regional features are provided the focus lies on object detection. A task very similar to this paper though using different features is presented in [2]. Also similar is [10] using Gabor features [11] on segmented satellite images which provide optimal conditions for texture analysis as they barely suffer from perspective distortion. The idea of this study is to ignore the perspective distortion and reduce the problem of identifying image regions to a problem of identifying textures. The publication closest to this work is [12]. It also faces the task of concept similarity measures in an even larger scale and therefore affiliation estimation. The features used are two-dimensional hidden Markov models.

- building
- foliage
- mountain
- person
- road
- sailing boat
- sand
- sea
- sky
- snow

TABLE I
THE PRE-DEFINED CONCEPTS TO BE FOUND IN IMAGES

[1] http://wang.ist.psu.edu/docs/related.shtml
[2] http://ganymed.imib.rwth-aachen.de/irma/index_en.php

III. Design

A set of 922 RGB-images with a resolution of 800 by 600 pixels was provided to represent ”holiday images”. They originate in the available Database of the K-Space Project[3]. To avoid an *over-fitting* of the (to be generated) concept descriptors the data was separated into three subsets. One for the training of the concept classifiers, a second one to validate parameter changes in the algorithm and a third one to test the final classification performance. For the training- and validation-subset 230 images were each randomly chosen out of the provided images. The remaining 462 photos were used as test-subset.

A. Concept Training

The aim of this study is to enable a comparison of image regions against a set of pre-defined concepts using WSDs. Therefore a WSD-representation of the concepts must be developed. As a WSD is a feature vector of invariant length (for identical image resolutions) it was chosen to represent each of the features by mean and standard-deviation. To generate the concept-descriptors the following steps were taken:

1) Compute the WSDs of each image region within the training-set and select one representative per region.
2) Group the representative WSDs of all the images within the training-set by their concept label using a manually annotated affiliation list.
3) Calculate the mean and standard-deviation of each feature value of a WSD over the given concept.
4) Store mean and standard-deviation as representative for the corresponding concept in a concept descriptor.

A concept is now defined by a label (e.g. sailing-boat) and a corresponding descriptor which is a mean and standard-deviation value for each computable feature value.

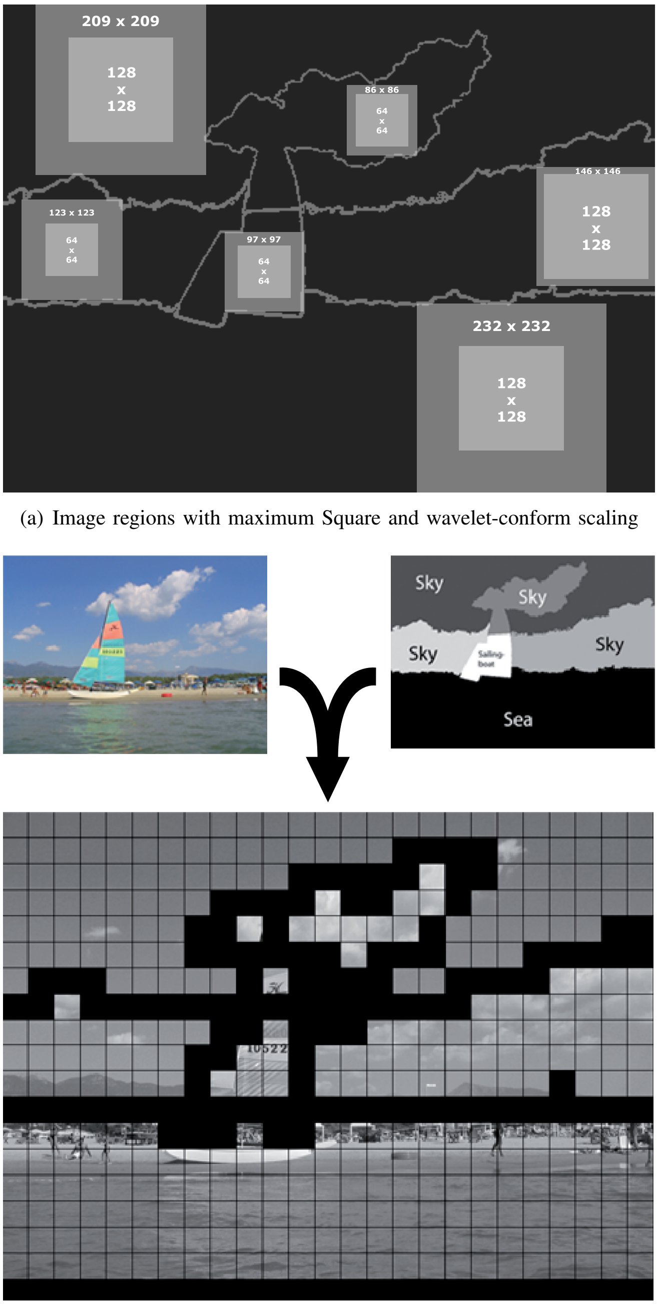
B. Modalities & Validation

When dealing with texture data there’s usually no need to provide color information as it multiplies the processing time at least by the number of color-channels used. There is also the problem that color information is strongly dependant on the environment’s illumination. However color channels can include patterns that supply additional information of the structural character. During the validation it showed that e.g. the concepts snow and sand tend to be very similar in their structural consistency. Here a separate analysis of the color channels can add additional precision to the results. Therefore three different color-spaces were tested - gray-scale, RGB, and YC$_r$C$_b$. The gray-scales are obviously representing structure-only information. RGB was chosen in regard to the simplicity of the approach as the available images are provided in this format and it is to be expected that most of the sources of ”holiday images” (e.g. cameras) have the same output. In case of satisfying results without color-space transformations the run-time performance would benefit. YC$_r$C$_b$ was chosen as proposed in [1] with the argument of describing structural information in the Y-component and the distribution of color information in the C$_r$- and C$_b$-components.

When working with wavelet transformations the image or region to be processed must fulfill the conditions

$$(x = y) \wedge (x = 2^n : n \in \mathbb{N}_{>0}) \quad (1)$$

where x and y define the pixel resolution. As image regions mostly don’t fit these requirements it is necessary to choose a representative for every region in order to compute the WSD-features. To meet these conditions two approaches were tested.



(a) Image regions with maximum Square and wavelet-conform scaling

(b) Determination of valid patches using a 32 by 32 pixels grid

Fig. 1. maximum Square 1(a) and grid overlay 1(b) feature selection methods

[3] http://kspace.qmul.net:8080/kspace/

a) maximum square: The first approach finds the largest square fitting in an image region and scales it down to the next valid wavelet-conform resolution as shown in fig. 1(a). Then the WSD was computed stopping after the third level of the wavelet transformation (see IV-A) in order to assure a common vector size for later comparison. This limits the patch resolution from 8^2 up to 512^2 pixels using 800 by 600 pixel images.

b) grid overlay: The second approach uses a square grid overlay to create a pattern with wavelet-conform patches. The size of the grid's patches was set to 32^2 pixels. When using $\mathrm{YC}_r\mathrm{C}_b$ color-space a resolution of 8^2 pixels was additionally tested. Only patches lying completely within a single region were considered for the WSD computation. The process of finding out the valid patches ist depicted in figure 1(b). To determine a single representative for an image region different methods were used:

- *first patch:*
  when processing the image, the first found patch of an image region was selected as representative (dummy implementation).
- *mean value:*
  the WSD of each patch was computed. All WSD features of an image region were merged into a representative WSD for the image region using mean for every feature (per filter and level).
- *maximum affiliation:*
  the WSD of each patch was computed. Then they were compared to each concept resulting in 10 affiliation values[4] per patch. The patch with the largest affiliation value[4] was selected as representative.
- *largest difference:*
  similar to *max. affiliation*. Though the representative was selected by the largest difference of the highest two affiliation values[4] per patch.

Table II lists the modes and combinations the validation was performed in. The rows show the representative selection method, the columns show the colorspace. Content is the used patch size of the specific method. The *maximum square* entry refers to the first selection approach and therefore doesn't use a static patch size. All parts of the algorithm were validated against the validation-set of images concerning error handling and parameter optimization.

	gray-scales	RGB	$\mathrm{YC}_r\mathrm{C}_b$
first patch	32^2	-	-
mean value		32^2	8^2 and 32^2
maximum affiliation			
largest difference			
maximum Square	8^2 to 512^2	-	-

TABLE II
VALIDATION PERFORMED REGARDING COLOR-SPACE AND REGION REPRESENTATIVE SELECTION METHOD. THE CONTENTS SHOW THE APPLIED PATCH RESOLUTION OF THE GRID OVERLAY IN PIXELS

C. Algorithm Testing

In order to avoid bias due to overtraining of the algorithms on a specific set all presented results in this paper were created using the images of the test-set with the unaltered parameters or fixes created during the validation process.

IV. Implementation

The implementation of this study covers four partial tasks.

- create the wavelet-transformation for the input images
- compute the WSD out of the wavelet-transformed images
- describe any pre-defined concept using WSDs
- a *concept-to-WSD* comparison method must be developed

As the descriptor for the WSD values was decided to be the mean and standard-deviation over the computed features in the training set, there won't be a more detailed description for point three. The implementation of the wavelet-transformation, the WSD computation and the selected comparison method are presented in the following.

A. The Wavelet Standard Deviation Descriptor

The wavelet-transformation computes the frequencies within different levels of a signal. In case of image processing it is interpreted as the frequencies of a single channel's discrete values within the input image - most likely gray-scale images. The implementation uses the Haar wavelet [14] following the proposal of [1]. It results in a transformed image like the example in figure 2. Every level of the wavelet transformation a partial HL-, LH-, HH- and LL-image is created where H is a high-pass filter and L a low-pass filter. The partial images are created following:

1) Apply the horizontal filter to the input image (e.g. H).
2) Eliminate every second column.
3) Apply the vertical filter (e.g. L).
4) Eliminate every second row.
5) In case of HL-, LH- and HH-images store the result in the specific position shown in figure 2 (e.g. top-right for HL). in case of an LL-image compute the next wavelet-level using the LL-image as input.
6) Repeat this process until the desired wavelet-level is reached or the LL-image has a resolution of 1 by 1 pixels.

The *wavelet standard deviation descriptor* (WSD,[1]) is a vector that uses the weighted standard deviation of each partial HL-, LH- and HH-image as feature values. The weighted standard-deviation of the LL-image and its mean are used as final feature values in the vector. The *maximum Square* approach uses a depth of three wavelet-levels as its WSD. In the *grid overlay* approach all level-features were calculated and used as WSD. In the second case the size of the feature vector is therefore always:

$$\varepsilon = 3 \cdot \#_k + 2 \tag{2}$$

Where $\#_k$ is the number of computed levels.

[4] described in IV-B

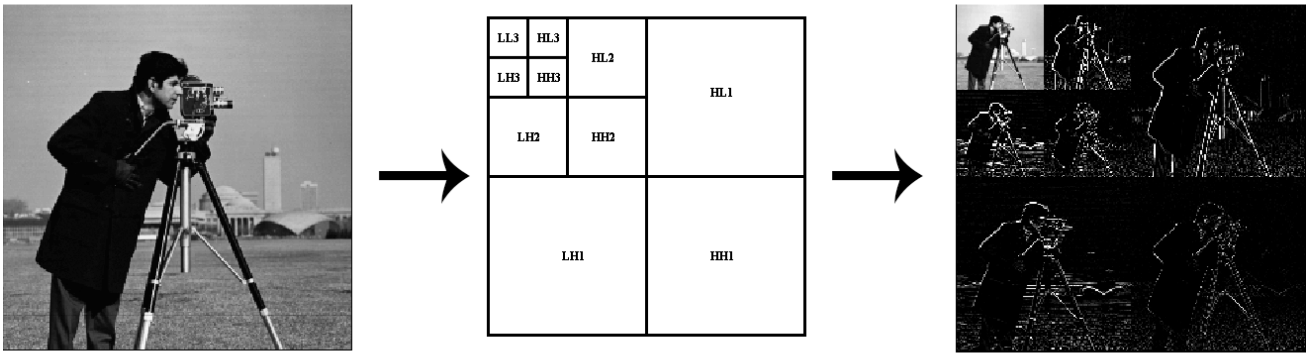


Fig. 2. Wavelet transformation with three levels on an image by [13]. The intensities in the LL-image on the right were normalized for visualization purposes.

The WSD vector is then defined as:

$$WSD = \{\frac{\sigma(LH_k)}{2^{k-1}}, \frac{\sigma(HL_k)}{2^{k-1}}, \frac{\sigma(HH_k)}{2^{k-1}}, \ldots, \frac{\sigma(A)}{2^{k-1}}, \mu(A)\} \tag{3}$$

$LH_k, HL_k, HH_k =$ the partial image of the k-th level
$\sigma(image) =$ standard deviation of the image
$\mu(image) =$ mean value of the image
$k =$ the index of the specific level $[1..(\varepsilon - 2)/3]$
A = the approximation image (the lowest LL-image)

B. Comparison Method

In [1] a similarity measure for WSDs is offered based on the weighted difference of the corresponding features. It was decided to implement a probabilistic measure similar to the offered approach. As after section III-A each feature of a concept is represented by mean μ and standard-deviation σ, it is possible to compute a relative similarity ρ for each feature f of a WSD using the probability distribution:

$$\rho_k = exp(-\frac{(f_k - \mu)^2}{2\sigma^2}), k \in [1..\varepsilon] \tag{4}$$

The final concept affiliation was tested with two different aggregation modes - the sum of the single values and their product.

$$\gamma_{sum} = \Sigma_1^{\varepsilon} \rho_k \tag{5}$$

$$\gamma_{product} = \Pi_1^{\varepsilon} \rho_k \tag{6}$$

V. Results

Within the 462 images of the test-set a manual annotation was provided for 2960 of the contained regions. Any result presented in this paper is referencing to this number of regions providing ground truth. Applying the grid presented in section III-B with a patch size of 32^2 pixels results in 2521 classifiable regions. The remainder of the regions didn't fit the grid in so far as none of its patches were completely located within its boundaries and therefore no features could be extracted. A patch size of 8^2 pixels could cover each region. Anyway these circumstances make it necessary to distinguish between two types of results. First there is the *feature classification rate* (feature rate) which represents the classification performance only in regard to all computable features. And second there is the *task classification rate* (task rate) which shows the overall classification performance treating ignored regions as falsely classified.

The computed classification rates are presented in Table III. It shows that a region representative is best selected by creating a mean WSD over all the region's patches. The feature aggregation to be favored is summing up the feature values (compare equation 5). Using color information doesn't strongly influence the classification rates though using RGB shows slight advantages. Applying the smaller sized grid improves the results by less than one percent when comparing the best classification rates of each size. Instead the algorithm's run time increases a lot as there are 16 times as many features to be computed, merged and compared. The *maximum squares* approach delivered results worse than the task rates of a 32^2 pixel sized grid, both using gray-scale converted images.

It showed up that none of the patches provided any data in the partial images after the fourth level of the wavelet-transformation. And also the fourth level only contained data in a very few cases. This was not tested with patches larger than 32^2 pixels but the computation speed of implementations as used in this study can be enhanced when stopping feature computation after the third wavelet level. Table II includes an entry for a *first patch* method on 32^2 pixel gray-scale patches. This method was only implemented as an early placeholder for better founded methods and is therefore not listed in the results table.

colorspace & patch size	patch selection	feature aggregation	regions		feature rate			task rate		
			total	validated	1st	2nd	3rd	1st	2nd	3rd
GRAY max. squares	first patch	$\sum$	2960	all	identical to task rate			32.20	44.46	55.20
		$\prod$						23.41	31.76	41.39
YCrCb 8x8	maximum affiliation	$\sum$						19.06	31.38	35.94
		$\prod$						14.64	28.26	34.28
	mean value	$\sum$						**36.45**	**53.77**	61.88
		$\prod$						32.25	51.88	59.57
	largest difference	$\sum$						4.78	15.22	21.96
		$\prod$						5.36	16.32	22.75
YCrCb 32x32	maximum affiliation	$\sum$		2521	40.16	58.54	59.89	34.20	49.86	51.01
		$\prod$			6.13	14.04	25.61	5.22	11.96	21.81
	mean value	$\sum$			38.29	**62.28**	**76.74**	32.61	53.04	**65.36**
		$\prod$			5.27	16.51	25.01	4.49	14.06	21.30
	largest difference	$\sum$			23.65	41.94	57.60	20.14	35.72	49.06
		$\prod$			6.13	14.12	25.61	5.22	12.03	21.81
RGB 32x32	maximum affiliation	$\sum$			37.01	52.33	70.03	31.52	44.57	59.64
		$\prod$			8.68	15.31	26.97	7.39	13.04	22.97
	mean value	$\sum$			**42.12**	60.49	75.81	35.87	51.52	64.57
		$\prod$			7.91	16.25	26.21	6.74	13.84	22.32
	largest difference	$\sum$			29.61	45.44	58.87	25.22	38.70	50.14
		$\prod$			8.85	15.15	26.97	7.54	12.90	22.97
GRAY 32x32	maximum affiliation	$\sum$			35.91	51.81	68.83	30.58	44.13	58.62
		$\prod$			7.66	15.57	26.47	6.52	13.26	22.54
	mean value	$\sum$			39.57	59.39	75.04	33.70	50.58	63.91
		$\prod$			6.55	15.40	25.69	5.58	13.12	21.88
	largest difference	$\sum$			25.61	41.86	55.48	21.81	35.65	47.25
		$\prod$			7.49	15.49	26.29	6.38	13.19	22.39

- *feature rate* represents the classification performance in regard to all validated regions
- *task rate* represents the classification performance in regard to the total number of regions
- 1st / 2nd / 3rd ≙ searched concept is contained in best / best two / best three affiliation result(s)
- all classification rates are percentage values, bold text shows the column's maximum
- $\sum$, $\prod$ reference to equations 5 and 6

TABLE III
CLASSIFICATION RESULTS

VI. CONCLUSION

This paper is based on the wavelet standard-deviation descriptor [1] that uses texture images from the Brodatz texture database [15] containing 1856 samples. Their algorithm reaches an average classification rate of up to 70.30% when used in an image retrieval scenario. Compared to the extended classification approach in this paper the maximum task rate of 36.45% is to be rated rather moderate. The maximum feature rate of 42.12% also shows that a classification system (e.g. an automated image tagging system) can't solely be based on this algorithm. Nevertheless in 76.74% of the cases the correct concept can be found within the best three affiliated concepts out of ten. This can be used as a weighting factor for algorithms dealing with similar tasks. The results show two major problems when using texture descriptors on segmented images of natural scenarios. First the shape or size of the segments can strongly counter-act the requirements of the texture descriptor - in case of wavelets this is the square-sized base shape. And second texture patterns in natural images underly a perspective distortion as you most likely don't watch top-down onto the surface (like in satellite photography). The extension of the original approach brings up one more issue. Instead of comparing two images directly, an image is compared to a trained descriptor representing a group of texture patterns. This affects the capabilities of the algorithm as it adds additional noise to the results. Considering those problems the results appear sufficient as base for upcoming studies and the presented knowledge gaps require further investigation. The following paragraphs propose approaches to be validated in order to improve the classification performance.

A. Segment WSD selection

The determination of a segment's WSD can be dealt with in very different ways. The grid-based process used in this paper (see section IV-B) was chosen to make sure the input for the wavelet-transformation has always the same resolution. Another aspect was a predictable computation duration. However this demands for a selection or aggregation of data as most likely many WSDs can be computed for a single segment. The larger the grid's patch size is compared to the segment size the more data is lost in the border regions of the segments. It also increases the possibility to find no single patch being completely located within the segment. The smaller the grid's patch size is chosen the less representing the data found in a single patch becomes. This also increases the computation time a lot as much more wavelet-transformations and comparisons to the concepts must be performed per segment. Upcoming studies can therefore focus on the problem of finding an optimal patch size depending on the image resolution, the segment size and shape.

The selection or creation of a representative patch currently favors the mean value methods presented in section III-B. A method yet unevaluated would be to count the appearances the patches' concept affiliations favor a specific concept within a single segment. Saying a segment consisting of 10 patches favors 2 times "sand", 3 times "water" and 5 times "snow". Then you can select the "snow"-patch with the highest affiliation value to the concept snow as representative. An extension to this idea would be to connect positioning values to a patch's concept affiliation values. Saying the highest value scores 10 points (when having 10 concepts), the next highest value scores 9 points and so on. In the end you sum up the concept affiliation scores and pick the concept with the highest score as representative.

Another Idea that was tested is the *maximum square* approach. This approach increases the computation speed a lot as only one feature vector has to be calculated as soon as the maximum square area was identified. Unfortunately the results show a worse performance than a grid sized approach on gray-scales. An even more sophisticated approach would be to identify subregions within the image segments that are oriented almost planar to the camera's viewpoint. Those subregions provide data that is just slightly altered due to perspective distortion. You can also think of finding a segment's world plane in order to create a planar representation through perspective back projection (e.g. an ocean layer or a wall). The assumption here is that the distortion a perspective back projection creates is smaller than the information value generated in terms of WSD similarity.

B. Similarity Measure

The similarity measure presented in this paper is based on a probabilistic approach using mean and standard deviation to create a similarity value for each of the feature's levels. Finally the similarities of each level were either summed up or multiplied to create a value expressing the patch's affiliation to a concept.

In literature there are a lot more similarity measures offered for vector- and histogram-like comparisons. Implementations using (e.g.) the vector angle, a support vector machine, the earth mover's distance or the kullback-leibler divergence might improve the results a lot.

C. Usage of color values

The implementation as proposed in [1] currently only uses the standard deviation of each color channel as color information. This means only the structural character of the color is taken into account when creating a concept representation. A more promising approach would be to use the color intensities and a comparison method based on their mean value as an additional factor in the process of calculating a concept affiliation. This could be realized e.g. using the HSV or Lab color space. A WSD-conform vector on the S,V or L channels is created to represent the structure of a segment and additionally the H or a and b channels create a similar vector including the color mean values. This should at least improve distinguishing concepts like snow, sand or sea from each other.

REFERENCES

[1] Sitaram Bhagavathy and Kapil Chhabra, "A wavelet-based image retrieval system," University of California, Santa Barbara, Tech. Rep., an ECE 278A Project Report.

[2] C. Carson, M. Thomas, S. Belongie, J. Hellerstein, and J. Malik, "Blobworld: a system for region-based image indexing and retrieval," Berkeley, CA, USA, Tech. Rep., 1999. [Online]. Available: http://portal.acm.org/citation.cfm?id=893714

[3] T. Deselaers, D. Keysers, and H. Ney, "Features for image retrieval - a quantitative comparison," in *In DAGM 2004, Pattern Recognition, 26th DAGM Symposium*, 2004, pp. 228–236.

[4] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz, "Efficient and effective querying by image content," *J. Intell. Inf. Syst.*, vol. 3, pp. 231–262, July 1994. [Online]. Available: http://dx.doi.org/10.1007/BF00962238

[5] D. ing Sven Siggelkow, D. Prof, D. T. Ottmann, P. Dr, T. Ottmann, B. Haasdonk, L. Bergen, O. Ronneberger, C. B. S. Utcke, and S. Siggelkow, "Feature histograms for content-based image retrieval," 2002.

[6] C. Palm, D. Keysers, T. Lehmann, and K. Spitzer, "Gabor filtering of complex hue/saturation images for color texture classification," 2000.

[7] H. Tamura, T. Mori, and T. Yamawaki, "Textural features corresponding to visual perception," vol. 8, pp. 460–473, June 1978.

[8] T. Deselaers, "Features for image retrieval," December 2003, diploma Thesis.

[9] J. Z. Wang, J. Li, and G. Wiederhold, "Simplicity: Semantics-sensitive integrated matching for picture libraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 947–963, 2001.

[10] B. Manjunath and W. Ma, "Browsing large satellite and aerial photographs," 1996, pp. II: 765–768.

[11] ——, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 837–842, 1996.

[12] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 25, no. 9, September 2003.

[13] Michael Clemens, "Wavelet tutorial," http://nt.eit.uni-kl.de/wavelet/dwt_2d.html (11. Mars 2010).

[14] A. Haar, "Zur theorie der orthogonalen funktionssysteme," *Mathematische Annalen*, vol. 69, no. 3, pp. 331–371, 1910.

[15] P. Brodatz, *Textures: A Photographic Album for Artists and Designers*. Dover Publications, 1999.