Objectionable Image Detection by ASSOM Competition

Grégoire Lefebvre, Huicheng Zheng, Christophe Laurent

Objectionable Image Detection by ASSOM Competition ⋆

Grégoire Lefebvre[1], Huicheng Zheng[2], and Christophe Laurent[1]

[1] France Telecom R&D
4, Rue du Clos Courtel
35512 Cesson Sévigné – France
{gregoire.lefebvre,christophe2.laurent}@francetelecom.com

[2] Trinity College Dublin
College Green
Dublin 2 – Ireland
zhengh@tcd.ie

Abstract. This article presents a method aiming at filtering objectionable image contents. This kind of problem is very similar to object recognition and image classification. In this paper, we propose to use Adaptive-Subspace Self-Organizing Maps (ASSOM) to generate invariant pornographic features. To reach this goal, we construct local signatures associated to salient patches according to adult and benign databases. Then, we feed these vectors into each specialized ASSOM neural network. At the end of the learning step, each neural unit is tuned to a particular local signature prototype. Thus, each input image generates two neural maps that can be represented by two activation vectors. A supervised learning is finally done by a Normalized Radial Basis Function (NRBF) network to decide the image category. This scheme offers very promising results for image classification with a percentage of 87.8% of correct classification rates.

1 Introduction

In many computer vision applications such as multimedia data mining, pattern recognition, image retrieval, etc., evaluating image content is fundamental. Recognizing harmful images is very challenging in content-based filtering systems.

In the state-of-the-art, different approaches are proposed, focusing on skin color detection. Forsyth et al. use geometric constraints for detecting naked people [1] by reconstructing the human anatomic structure from skin areas. The WIPE system [2] combines Daubechies wavelets, moment analysis and histogram indexing to provide semantically meaningful feature vector matching. Based on the discrete probability distributions obtained from skin and non-skin histograms, Jones and Rehg [3] filter images according to their skin pixel statistics. Other studies [4, 5] propose a pornographic image detection system based on skin detection and Multi-Layer Perceptron (MLP) classification.

To avoid inherent skin detection problems, as illumination variations, background interferences, multiple figures, etc., we consider objectionable image filtering as image classification problem in two categories: adult and benign. Image

⋆ This work was carried out during the tenure of a MUSCLE Internal fellowship.

classification consists in partitioning the input image space into a number of regions separated by decision surfaces and labeled by image classes.

For a given image $\mathcal{I}$, the ultimate goal is to search for a function $f(\cdot) \rightarrow \mathcal{J}_{\mathcal{I}}$, $\mathbb{I} \rightarrow \mathbb{J}$, where $\mathbb{I}$ is the image space and $\mathbb{J}$ the image label space. However, due to the extremely high dimension of an ordinary image, a direct search of the optimum function $f(\cdot)$ in the original image space $\mathbb{I}$ would generally not be possible.

From this observation, we try to describe the image $\mathcal{I}$ in a more compact way to reduce the dimension of data and obtain the most discriminant informations. In this purpose, we are interested in generating invariant-feature descriptors by using ASSOM neural networks [6].

Adaptive-Subspace Self-Organizing Map (ASSOM) is basically a combination of a subspace method and a competitive selection and cooperative learning as in the traditional SOM [6]. The single weight vectors at map units in SOM are replaced by sets of basis vectors that span some linear subspaces in ASSOM. A long-standing difficulty in the design of feature filters is the variation of input patterns due to typical transformations such as translation, rotation and scaling. By setting filters to correspond to pattern subspaces, some transformation groups can be taken into account automatically.

The input to an ASSOM network is called "episode", which is a sequence of pattern vectors that spans some linear subspace. This sequence is contructed by applying rotation, translation and scaling to the original local signatures. By learning the episode as a whole, ASSOM is able to capture the transformation kernels coded in the episode.

To construct these episodes, we focus our attention on regions of interest (ROI) in the images. Based on some psycho-visual experiments [7], human vision system executes saccadic eye movements between salient locations to capture image content. Likewise, Tversky studies [8] showed that when we compare two images, we detect common and distinct concepts between these regions.

Our method tries to reproduce this extraction and distinction concept with a codebook learning strategy based on ASSOM algorithm. We firstly search salient locations in the images to be compared. Local visual features are then extracted from salient regions and projected onto a set of ASSOM-based learned visual prototypes, resulting in activation vectors.

Finally, we use these activation features to classify the image in adult or benign category with a NRBF neural network.

This method has been experimented for an adult content filtering method where the correct classification rate reaches 87.8%. The database is composed of $1,110$ adult images and $1,200$ benign images downloaded from Internet. The second category, known to be the rest of the world, is mainly constituted of landscapes, portraits and life scenes.

This paper is organized as follows: In Section 2, we first present our image classification scheme based on ASSOM learning from ROI descriptions. Then, Section 3 contributes to some experimental results on the proposed classification schemes. And finally, conclusions are discussed in Section 4.

2 Image Classification Based on ASSOM Learning and Salient Regions of Interest

2.1 Multi-ASSOM Scheme (MAS).

As outlined in [9], a classification scheme is generally composed of three main steps : pre-processing, feature extraction and classification. In this paper, we mainly focus our attention on the two first items, the last being performed by a NRBF neural network.

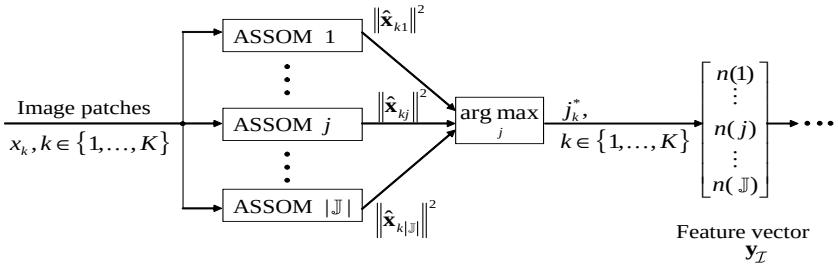

Fig. 1. The construction of the feature vector $\mathbf{y}_{\mathcal{I}}$ from patches of the image $\mathcal{I}$ in MAS. $|\mathbb{J}|$ ASSOMs compete on these patches and generate a sequence of winning ASSOM index j_k^*, $k \in \{1, 2, \ldots, K\}$. $n(\cdot)$ counts the number of winning times for each ASSOM. The vector $[n(1), \ldots, n(j), \ldots, n(|\mathbb{J}|)]^{\mathrm{T}}$ forms the final feature $\mathbf{y}_{\mathcal{I}}$, which is sent to the NRBF neural network

Our system architecture designs an ASSOM for each category, producing specific ASSOM units for different categories of image patches. This idea was explored in [10] in the recognition of handwritten digits and produced promising results. But in their case, the image size is very small (25×20 pixels), permitting a direct learning through ASSOM. In their work, 10 ASSOMs are employed, one trained for each category of handwritten digits. For digit classification, a test digit is sent simultaneously to all the 10 ASSOMs, which output 10 error values. The ASSOM with the least reconstruction error determines the digit category. An obvious deficiency here is that there is no interaction between the different ASSOMs during the learning phase. An ASSOM learns the features of its own category, however it does not learn to distinguish features of other categories. The optimum decision surface is thus not guaranteed.

In our context, the dealt images have much larger sizes. So, we decide to use a local approach by extracting round image patches at salient locations. RGB patch informations are directly used to describe the local visual features.

Similar strategies have been developped in [11, 12]. The principal differences are that the bag-of-keypoints representation is built from a K-means quantization and the classification is made by a Support Vector Machine(SVM).

Here, $|\mathbb{J}|$ ASSOMs are trained on these local descriptions, category by category. $\mathbb{J}$ denotes the number of ASSOM. For filtering use, two ASSOM neural networks are built for adult and benign classes.

To construct the feature vector $\mathbf{y}_{\mathcal{I}}$ for the final NRBF classification of the image $\mathcal{I}$ (See Figure 1), we operate as follows :

– For each patch $\mathbf{x}_k$, the $|\mathbb{J}|$ ASSOMs compete on it. The jth ASSOM produces an output $\|\hat{\mathbf{x}}_{kj}\|^2$ defined by:

$$\|\hat{\mathbf{x}}_{kj}\|^2 = \max_{i \in I_j} \|\hat{\mathbf{x}}_{k\mathcal{L}_i}\|^2 \ , \tag{1}$$

where I_j is the set of indices of the modules in the jth ASSOM. In words, $\|\hat{\mathbf{x}}_{kj}\|^2$ is the maximum value of the square of the orthogonal projection of the patch $\mathbf{x}_k$ on the subspaces of the modules in the jth ASSOM. The j_k^*th ASSOM with the maximal output wins that patch:

$$j_k^* = \underset{j \in \{1,2,\ldots,|\mathbb{J}|\}}{\arg\max} \|\hat{\mathbf{x}}_{kj}\|^2 \ . \tag{2}$$

– A counter $n(j_k^*)$ corresponding to this ASSOM network is accordingly increased by 1. When all the patches of the image $\mathcal{I}$ have been presented, the array of ASSOMs produce $|\mathbb{J}|$ counters of the patches won by the respective ASSOM. The feature vector is defined by:

$$\mathbf{y}_{\mathcal{I}} = [n(1), \ldots, n(j), \ldots, n(|\mathbb{J}|)]^{\mathrm{T}} \ , \tag{3}$$

where the components are:

$$\forall j \in \{1, 2, \ldots, |\mathbb{J}|\}, \quad n(j) = \sum_{k \in \{1,2,\ldots,K\}} \delta(j_k^*, j) \ . \tag{4}$$

$\delta(a, b)$ is the pulse function that takes the value 1 when $a = b$ and 0 otherwise.

2.2 Wavelet-Based Salient Point Detection

According to the active vision mechanisms, the goal of salient point detectors is to find perceptually relevant image locations. Many detectors have been proposed in the literature [13–15]. The Harris detector [14] aims at locating salient zones on corners by searching for the maxima of a function based on the local autocorrelation matrix of the signal. The detector in [15] proposes to locate salient points in high contrasted area. The salient point detector in [13] uses a wavelet analysis to find pixels on sharp region boundaries.

Working with wavelets in our previous work [13] is justified by the consideration of the human visual system for which multi-resolution, orientation and frequency analysis is of prime importance. In order to extract the salient points, a wavelet transform is firstly performed on the grayscale image. The obtained wavelet coefficients are represented by zerotrees as introduced by Shapiro [16].

This tree is then scanned at a first time from leaves to the root to compute the saliency value at each node. A second scanning occurs in order to determine the salient path from the root to the locations on the original image, where the raw salient points are located. The salient points are listed in order and a threshold τ, $0 < \tau \leq 1$ is set to select the most salient points. By detecting salient points from luminance information only, the points located on boundaries of highlights or shadows are apt to be detected as salient. To remove false salient points caused by lumination conditions, a gradient image is built by using the color invariants proposed by Geusebroek et al. [17].

This salient point detector reaches photometric invariance by combining the detection step with a recently proposed color invariance method [17]. Experimental results in [13] show that the detected points are located on perceptually relevant image areas. Based on the detected salient points, the authors went further to design the salient signature that combines a color histogram with a texture measure. The proposed salient point detector and salient signature are applied to a contented-based retrieval system and the results are quite promising.

In this paper, we will combine the salient point detectors of our previous work with the ASSOM feature selector. Consequently, the ASSOM networks can be trained on small image patches centered on these salient points.

2.3 ASSOM Learning

As mentioned in the introduction, ASSOM is basically a combination of a subspace method and a competitive selection and cooperative learning as in the traditional SOM. ASSOM differs from other subspace methods by permitting to generate a set of topologically-ordered subspaces. That is to say, two units that are close in the map will represent two feature subspaces closed in the total feature space. In ASSOM, the unit is composed of several basic vectors that expand together a linear subspace. This unit is called "module" in an ASSOM neural network. This method aims to learn data features, without assuming any prior mathematical forms of their representation, such as Gabor or wavelet transforms, which are frequently encountered in the traditional image analysis and pattern recognition techniques [6]. In other words, the forms of the filter functions are learned directly from the data.

The input to ASSOM is a group of vectors, called "episode". The vectors in each episode are supposed to be close up to affine transformations.
There are mainly two phases in a learning process of ASSOM:

1. For an input episode, locate the winning subspace from ASSOM modules ;
2. Adjust the winning subspace and its neighbor modules in order to better represent the input episode.

For a linear subspace $\mathcal{L}$ of dimensionality H, one can find a set of basis vectors $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_H\}$, such that every vector in $\mathcal{L}$ can be represented by a linear combination of these basis vectors. Such sets of basis vectors are not unique, however they are equivalent in the sense that they expand exactly the

same subspace. For convenience of mathematical measures, the basis vectors are orthonormalized by the Gram-Schmidt process.

The orthogonal projection of an arbitrary vector $\mathbf{x}$ on the subspace $\mathcal{L}$, notated as $\hat{\mathbf{x}}_{\mathcal{L}}$, is a linear combination of its orthogonal projections on the individual basis vectors, and can be computed by :

$$\hat{\mathbf{x}}_{\mathcal{L}} = \sum_{h=1}^{H} (\mathbf{x}^{\mathrm{T}} \mathbf{b}_h) \mathbf{b}_h. \tag{5}$$

If $\hat{\mathbf{x}}_{\mathcal{L}} = \mathbf{x}$, then $\mathbf{x}$ belongs to $\mathcal{L}$, else we can define the distance from $\mathbf{x}$ to $\mathcal{L}$ as $\|\tilde{\mathbf{x}}_{\mathcal{L}}\| = \|\mathbf{x} - \hat{\mathbf{x}}_{\mathcal{L}}\|$, by using the Euclidean norm. When several subspaces exist, the original space is separated from pattern zones and the decision surface between two subspaces, for example $\mathcal{L}_1$ and $\mathcal{L}_2$, is determined by those vectors $\mathbf{x}$ such that $\|\hat{\mathbf{x}}_{\mathcal{L}_1}\| = \|\hat{\mathbf{x}}_{\mathcal{L}_2}\|$. By comparing the distances of a vector to all the subspaces, we can assign this vector to the nearest subspace.

In Kohonen's realization of ASSOM, the subspace is represented by a two-layered neural architecture, as in Figure 2. The neurons in the first layer take the orthogonal projections $\mathbf{x}^{\mathrm{T}}\mathbf{b}_h$ of the input vector $\mathbf{x}$ on the individual basis vectors $\mathbf{b}_h$. The second layer is composed of a single quadratic neuron and makes the output square sum from the first layer neuron.

The output of the whole neural module is then $\|\hat{\mathbf{x}}_{\mathcal{L}}\|^2$, the square of the norm of the projection. It can be regarded as a measure of the degree of matching of the input vector $\mathbf{x}$ with the subspace $\mathcal{L}$ represented by the neural module. In the case of an episode, the distance should be calculated from the subspace of the vectors in the episode and that of the module, which are generally difficult to compute. Kohonen proposed another much easier but robust definition of subspace matching : the *energy* (sum of squares) of orthogonal input vector projections on the module subspace.

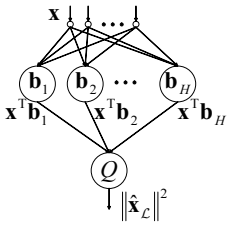

Fig. 2. Neural architecture of orthogonal projection of $\mathbf{x}$ on $\mathcal{L}$

Once the first phase occurred, the winning module with its neighbors adjust their subspaces to represent better the input subspace. A neighborhood function $h_c^{(i)}$ is defined on the rectangular or hexagonal lattice (See Figure 3), where c notates the index of the winning module and i the index of an arbitrary module

in the lattice. The neighborhood area defined by $h_c^{(i)}$ shrinks with the learning step. Through this cooperative learning, the map will end at a topologically-organized status, where nearby modules have similar subspaces.

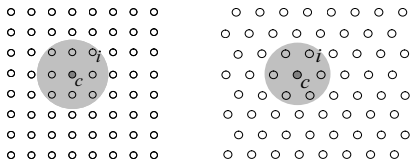

Fig. 3. Left: A rectangular topology. Right: A hexagonal topology. Each empty circle represents a neural module as shown in Fig. 2. The gray areas represent the neighborhood of the winning module indexed by c at a certain learning step.

The classical Kohonen's ASSOM learning algorithm works as follows:
For the learning step t,

1. Feed the input episode $\mathbf{x}(s)$, $s \in S$, where S is the set of indices of vectors in the input episode. Locate the winning module indexed by c:
$$c = \arg\max_{i \in I} \sum_{s \in S} \|\hat{\mathbf{x}}_{\mathcal{L}_i}(s)\|^2, \tag{6}$$
where I is the set of indices of the neural modules in the ASSOM.
2. For each module i in the neighborhood of c, including c itself, and for each input vector $\mathbf{x}(s)$, $s \in S$, adjust the subspace $\mathcal{L}_i$ by updating the basis vectors $\mathbf{b}_h^{(i)}$, according to the following procedure:
   (a) Rotate each basis vector according to:
   $$\mathbf{b}_h^{(i)} = \mathbf{P}_c^{(i)}(\mathbf{x}, t)\mathbf{b}_h^{'(i)} \ . \tag{7}$$
   In this updating rule, $\mathbf{b}_h^{(i)}$ is the new basis vector after rotation and $\mathbf{b}_h^{'(i)}$ the old one. $\mathbf{P}_c^{(i)}(\mathbf{x}, t)$ is the rotation operator matrix, which is defined by:
   $$\mathbf{P}_c^{(i)}(\mathbf{x}, t) = \mathbf{I} + \lambda(t)h_c^{(i)}(t)\frac{\mathbf{x}(s)\mathbf{x}^{\mathrm{T}}(s)}{\|\hat{\mathbf{x}}_{\mathcal{L}_i}(s)\|\|\mathbf{x}(s)\|} \ , \tag{8}$$
   where $\mathbf{I}$ is the identity matrix, $\lambda(t)$ a learning-rate factor that decreases with the learning step t. $h_c^{(i)}(t)$ is the neighborhood function defined on the ASSOM lattice with the support area shrinking with t.
   (b) Dissipate the components $b_{hj}^{(i)}$ of the basis vectors $\mathbf{b}_h^{(i)}$ to improve the stability of the results [6]:
   $$b_{hj}^{\sim(i)} = \mathrm{sgn}(b_{hj}^{(i)})\max(0, |b_{hj}^{(i)}| - \varepsilon), \tag{9}$$
   where ε is the amount of dissipation, chosen proportional to the magnitude of the correction of the basis vectors.
   (c) Orthonormalize the basis vectors in module i.

3 Experiments

In our experiment, the MAS scheme is applied to adult image filtering. There are respectively 733 adult images and 733 benign images in the training set of the image database. The test database is composed of 377 adult images and 467 benign images. The training of each ASSOM for each category takes $T = 200,000$ epochs. The subspace dimension is set to $H = 4$ and the dimension of the ASSOM arrays is $N = 10 \times 10$. The radius of a image patch is $r = 14.5$ pixels and thus the dimension of the input vector for the ASSOM arrays is $1,971$. These parameters are chosen by experimental results on the training set.

For our experiments, we configure our ASSOM network with the following rules to reach good learning results in terms of accurate input data representation [6]:

- $\lambda(t) = \frac{T}{T+99t}$ forms a monotonically decreasing sequence;
- $h_c^{(i)}(t) = \begin{cases} 1, ||p_c - p_i|| < \mu(t) \\ 0, otherwise. \end{cases}$

  The Euclidian norm is chosen and p_i is the 2D location for the i^{th} neuron in the network. $\mu(t)$ specifies the width of the neighborhood decreasing linearly during time t from $\frac{\sqrt{2}}{2}N$ to 0.5 .

The classification performance on the training set is a true positive (TP) rate of 88.5% at a false positive (FP) rate of 13.2%. The TP rate is defined as the proportion of adult images that are correctly classified and the FP rate is the proportion of benign images that are incorrectly classified as adult. On the test set, the TP rate is 89.9% and the FP rate is 13.9%. The confusion matrix is presented in the Table 1. Some examples of correct and incorrect classification are then shown in Figure 4.

The performance is really promising, considering the value 0.9448 for the area under the curve (AUC), drawn in the Receiver Operating Characteristics (ROC) curve in Figure 5.

This multi-ASSOM architecture outperforms with 87.8% of correct classification rates a single ASSOM scheme (85.9%) and a single SOM scheme (78.35%)[1].

Thus, we can see that the competition between an adult ASSOM and a benign ASSOM creates more precise feature vectors for NRBF classification. And the ASSOM algorithm permits to extract more robust features than the basic SOM method.

It is also very interesting to see the filters generated for the adult images and the benign images in Figure 6. We can observe that for the adult images, one of the basis vectors exhibit a orange color tone, and the other one shows obvious orientations. Thus, each final subspace tries to represent the data feature structure. That's why, when an adult image is fed into our scheme the adult ASSOM is stronger activated as shown in Figure 6.

[1] In our experiment, the best configuration for a single ASSOM scheme is : $N = 10 \times 10, H = 4, r = 14.5$; and for a single SOM scheme : $N = 20 \times 20, r = 3.5$.

Table 1. Confusion matrix of the MAS classification system with a ASSOM array of dimension 10×10. In this table, A=Adult, B=Benign.

Classified as →	A	B
A	339	38
B	65	402

Fig. 4. The left column presents correctly classified test images. The right column shows examples for misclassification.

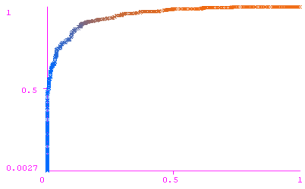

Fig. 5. ROC curve of MAS on classification of adult and benign test images. Horizontal axis represents the FP rate and vertical axis the TP rate.

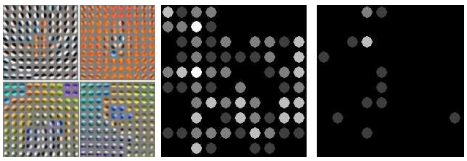

Fig. 6. (a)The feature filters generated for adult images and benign images. Top row: The filters generated for adult images. Bottom row: The filters generated for benign images. First column: The first basis vectors. Second column: The second basis vectors. The activation map represents the output energy for each module of adult ASSOM(b1) and benign ASSOM(b2) when an adult test image is proposed to MAS.

4 Conclusion

In this paper, we proposed an original classification system using directly patch information. Based on the three main properties of ASSOM - which are dimension reduction, topology preservation and invariant feature emergence - our scheme filters images in a competitive way. This solution implemented for content-based image filtering gives us very promising results. A further improvement could be to distinguish model portraits from adult images, because this case composes chiefly the false detection. To get efficiency, the adult class training can be achieved with an object-based approach in order to skip background interferences. Furthermore, MPEG7 descriptors may certainly make the performances better than the simple RGB informations.

References

1. Forsyth D.A., Fleck M.M.: Identifying nude pictures. IEEE WACV (1996) 103–108
2. Wang J.Z., Li J., Wiederhold G., Firschein O.: Classifying objectionable websites based on image content. IDMS (1998) 113–124
3. Jones M.J., Rehg J.M.: Statistical color models with application to skin detection. IJCV **46**(1) (2002) 81–96
4. Bosson A., Cawley G.C., Chan Y., Harvey R.: Non-retrieval: Blocking pornographic images. CIVR (2002) 50–60
5. Zheng H., Daoudi M., Jedynak B.: Blocking adult images based on statistical skin detection. ELCVIA **4**(2) (2004) 1–14
6. Kohonen T.: Self-Organizing Maps. Springer-Verlag, Berlin, Heidelberg, New York (2001)
7. Hoffman J.E., Subramanium B.: The role of visual attention in saccadic eye movements. Perception and Psychophysics **57** (1995) 787–795
8. Tversky A.: Features of similarity. Psychological Review **4**(84) (1977) 327–352
9. Duda R.O., Stork D.G., Hart P.E.: Pattern Classification. Wiley Interscience (2000)
10. Zhang, B., Fu, M., Yan, H., Jabri, M.A.: Handwritten digit recognition by adaptive-subspace self-organizing map (assom). IEEE Transactions on Neural Networks **4**(10) (1999) 939–945
11. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. ECCV (2004)
12. Quelhas, P., Monay, F., Odobez, J.M., Gatica-Perez, D., Tuytelaars, T., Gool, L.V.: Modeling scenes with local descriptors and latent aspects. ICCV (2005) 883–890
13. Laurent C., Laurent N., Maurizot M., Dorval T.: In depth analysis and evaluation of saliency-based color image indexing methods using wavelet salient features. Multimedia Tools and Application (2004)
14. Harris C., Stephens M.: A combined corner and edge detector. Proc. Fourth Alvey Vision Conf. (1988) 147–151
15. Bres S., Jolion J.M.: Detection of interest points for image indexation. In 3rd Int. Conf. on Visual Information Systems (1999) 427–434
16. Shapiro, J.: Embedded image coding using zerotrees of wavelet coefficients. IEEE Transactions on Signal Processing **12**(41) (1993) 3345–3462
17. Geusebroek, J.M., Boomgrad, R., Smeulders, W.M., Geerts, H.: Color invariance. IEEE Transactions on PAMI **12**(23) (2001) 1338–1350