A Weakly Supervised Approach for Semantic Image Indexing and Retrieval

Nicolas Maillot and Monique Thonnat

INRIA Sophia Antipolis - Orion Team,
2004 Route des lucioles - B.P. 93,
06902 Sophia Antipolis, France
{nicolas.maillot, monique.thonnat}@sophia.inria.fr

Abstract. This paper presents a new approach for building semantic image indexing and retrieval systems. Our approach is composed of four phases : (1) knowledge acquisition, (2) weakly-supervised learning, (3) indexing and (4) retrieval. Phase 1 is driven by a visual concept ontology which helps the expert to define low-level features useful to characterize object classes. Phase 2 uses acquired knowledge and image samples to learn the mapping between image data and visual concepts. Image indexing phase (phase 3) is fully automatic and produces semantic annotations of the images to index. The symbolic nature of querying enables user-friendly and fast retrieval (phase 4). We have applied our approach to the domain of transport vehicles (i.e. motorbikes, aircrafts, cars).

Keywords: Semantic-based retrieval; Learning in retrieval; Content analysis and understanding.

1 Introduction

This paper presents a new approach for building semantic image indexing and retrieval systems. We show how a priori knowledge provided by a domain expert can lead to an efficient semantic image indexing system. Our approach is composed of four phases : (1) a knowledge acquisition phase, (2) a weakly-supervised learning phase, (3) an indexing phase and (4) a retrieval phase.

This paper is structured as following. Section 2 gives an overview of existing semantic image indexing and retrieval approaches. Section 3 shows how the **domain knowledge acquisition** phase produces a hierarchy of classes described by visual concepts. Section 4 details the **weakly supervised learning** phase which consists of obtaining samples of the visual concepts used during knowledge acquisition for training a set of visual concept detectors. Section 5 is dedicated to the **semantic image indexing and retrieval** phases. **Indexing** uses the visual concept detectors trained during the weakly supervised image indexing phase. The **retrieval** phase is based on symbolic annotations computed during the semantic indexing phase and does not require any image processing capabilities. Section 6 is dedicated to results obtained on the problem of retrieval of images containing transport vehicles. We finally conclude and sketch future works in section 7.

W.-K. Leow et al. (Eds.): CIVR 2005, LNCS 3568, pp. 629–638, 2005.

2 Related Works

Image conceptual indexing and retrieval paradigm is now a topic of great interest.This stems from the limits of the query by example paradigm where image samples have to be provided : as explained in [1], one or several query image(s) cannot capture the conceptual essence of the user query.

Some techniques use manual annotations of images [2]. In this case, retrieval uses these annotations. Image processing is not used for indexing and retrieval.

In [3], querying is based on a logical composition of region templates. As explained by the authors, this approach is at an intermediate semantic level. One goal of the authors of this work is to reach a higher semantic level.

In [4], a statistical approach learns keywords describing images. A set of manually annotated images is used to enable learning. Due to the fact that no a priori knowledge is used, this approach often lead to semantically inconsistent image annotation. As explained by the authors of [4], a rule-based engine should be used to improve image interpretation consistency.

In [5], querying is based on an object ontology which defines the mapping between low level descriptors and intermediate level semantic notions. The system is used in two phases. Each concept (color, position, size, shape) of the proposed ontology is defined by the appropriate range of numerical values of the corresponding low level descriptors computed in image regions (e.g. luminance, hue). These generic constraints lead to coarse retrieval results. User feedback is then used to train support vector machines dedicated to constraint refinement. This approach relies on a cumbersome numerical descriptor database and does not propose a well defined formalism for high-level knowledge.

In [1], an image retrieval approach based on an extensible ontology is proposed. Querying is achieved by combining ontological concepts (e.g. size, location, color, semantic category). This combination is constrained by a grammar. Mapping between image data and concepts is based on supervised machine learning techniques (i.e. multi-layer perceptrons and radial basis networks).

A look on the state of the art shows that the community is trying to find a trade off between the amount of work needed to build image indexing and retrieval systems (e.g. supervised learning, manual annotation) and semantic richness. Our work also deals with this trade off and brings improvements on the state of the art. We propose a well formalized high-level knowledge (e.g. subsumption, part-whole and spatial relations) and we limit the amount of work needed to build image indexing and retrieval systems by using weakly supervised learning techniques.

3 Knowledge Acquisition and Formalization

First comes the knowledge acquisition phase which aims at capturing both high-level semantic categories and their visual description. More details can be found in [6]. This phase is driven by a visual concept ontology. As seen in fig. 1, knowledge acquisition consists of achieving the following tasks : domain taxonomy

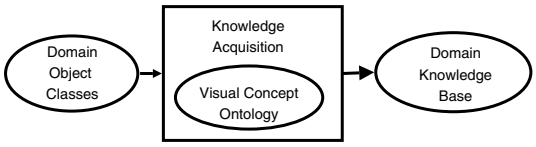

Fig. 1. Knowledge acquisition phase overview

acquisition (i.e. hierarchy of domain classes) and ontology driven visual description of domain classes which leads to a domain knowledge base.

The complete ontology is composed of 103 visual concepts (e.g. *Granulated* Texture, *Coarse* Texture, *Circular* Surface, *Dark*, *Elongated*, *Small*, *Circular*, *Pink*). The depth of the ontological tree is 8. This ontology is an extendible basis that can be specialized depending on the application domain. Numerical features are associated with visual concepts and define how visual concepts are computed on image data. Examples of numerical features associated with visual concept are : color coherence vectors [7] for visual concept *Hue*; co-occurence matrices [8] for visual concept *Pattern*; SIFT features [9] and MPEG-7 shape features [10] for visual concept *Geometry*.

Definition 1. *Let Θ be the set of all visual concepts. $\preceq_\Theta$ is a partial order between visual concepts. $\forall(C_i, C_j) \in \Theta^2, C_i \preceq_\Theta C_j$ means that C_i is a sub-concept of C_j.*

Definition 2. *Let Φ be the set of domain classes. For $\alpha \in \Phi$, $\mathcal{S}(\alpha) \subset \Phi$ is the set of subparts of α (i.e. subparts attribute).*

Definition 3. *Let $\mathcal{A} \subset \Theta$ be the set of domain class intrinsic attributes. $\mathcal{A}$ is a predefined subset of Θ. $\mathcal{A} = \{Geometry, Size, Orientation, Position, Hue, Brightness, Saturation, Repartition, Contrast, Pattern\}$. For a class $\alpha \in \Phi$, $\mathcal{A}_\alpha \subseteq \mathcal{A}$ is the set of attributes of α. $\preceq_\Phi$ is a partial order between domain classes (i.e. superclass attribute).*

Definition 4. *Let $a \in \mathcal{A}_\alpha$ be an attribute of $\alpha \in \Phi$. $\mathcal{V}_\alpha(a)$ is the set of possible values of a so that $\forall C \in \mathcal{V}_\alpha(a), C \preceq_\Theta a$ and $C \neq a$.*

Knowledge acquisition phase consists of defining Φ (i.e. the classes), $\preceq_\Phi$ (i.e. the class hierarchy), $\mathcal{S}(\alpha)$ (i.e. the subparts) and $\mathcal{V}_\alpha(a)$ (i.e. the visual description of domain classes). Φ, $\preceq_\Phi$ and $\mathcal{S}(\alpha)$ belong to domain knowledge. This knowledge is shared by the specialists of the domain. It is also independant of any vision layer and can be reused for other purposes. Defining $\mathcal{V}_\alpha(a)$ allows to reduce the semantic gap between expert knowledge and image level. As explained in the next section, this semantic gap is completely filled during a learning phase. Examples of classes are shown in table 1. This example results from a knowledge acquisition phase. For $\alpha = \{OutdoorScene\}$, $\mathcal{S}(\alpha) = \{Background, Object\}$. For $\alpha = \{Sky\}$, $\mathcal{A}_\alpha = \{Hue, Brightness, Pattern\}$ and the range of attribute Hue

Table 1. High level description of some domain classes. Attributes names are in **bold face**. Attribute possible values are in *italic*. Expert terminology is in SMALL CAPS

Domain Class	OUTDOORSCENE
SubParts:	
BACKGROUND	{SKY ASPHALT LANDSCAPE}
OBJECT	{AIRCRAFT CAR MOTORBIKE }
Relation Description :	
Centered:	{OBJECT}
Top:	{BACKGROUND}
Bottom:	{BACKGROUND}
Domain Class	SKY
ColorAttributes :	
Hue:	{*Blue Grey* }
Brightness:	{*Dark Bright*}
TextureAttributes :	
Pattern:	{*SmoothTexture*}
Domain Class	AIRCRAFT
SuperClass:	FLYINGOBJECT
SpatialAttributes :	
Geometry:	{*AircraftShape*}

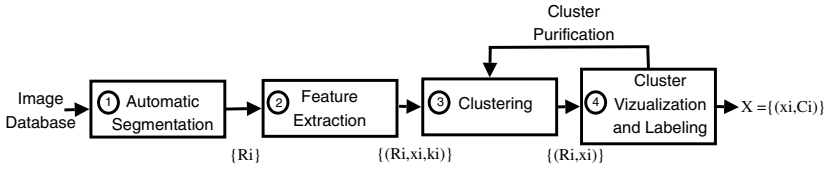

Fig. 2. From images to semantically labeled feature vectors. One execution of the sequence composed of steps (2),(3) and (4) corresponds to one visual concept of $\mathcal{A}$. Each visual concept contained in $\mathcal{A}$ is associated with different features. Depending on the considered visual concept of $\mathcal{A}$, feature extraction and clustering lead to different types of clusters (e.g. clusters resulting from regions of similar hue or of similar size)

is defined as $\mathcal{V}(Hue) = \{Blue, Grey\}$. The aquired knowledge base also contains the classes *AerialScene* and *RoadScene* which are subclasses of *OutdoorScene*. *AircraftShape* is domain specific and is a sub-concept of $PolygonalSurface$. This is the way to express that the geometry of an aircraft (e.g. sharp edges and corners) is a specific case of a polygonal surface.

As explained in [6], the proposed visual concept ontology stands as a meaningful user interface to a wide range of low-level image processing algorithms. A strong advantage of our approach is that improvements at the image processing level have no influence at the conceptual level.

4 Weakly Supervised Visual Concept Learning

In section 3, we have explained how the knowledge acquisition process leads to a set of domain classes described by visual concepts. One remaining and difficult issue is to fill the semantic gap between visual concepts and extracted low-level image data. This section aims at showing how this gap is filled by machine

learning techniques which lead to a set of visual concept detectors. Note that our goal is to obtain samples of visual concepts and not samples of domain classes. In other words, we simplify the problem by addressing it at an intermediate level of semantics. In [6], it is shown how region labeling by visual concepts is achieved manually. Manual segmentation and annotation of regions of interest by visual concepts was required. This tedious task is eased by clustering techniques.

Cluster labeling is divided into the following steps : automatic segmentation; feature extraction; clustering and cluster visualization and labeling (fig. 2).

(1) All the images of the image database are segmented into a set of regions $\{R_i\}$. Once the segmentation process is over, the sequence composed of steps (2),(3) and(4) is executed for each $a \in \mathcal{A}$ used during knowledge acquisition (i.e. $\exists \mathcal{A}_\alpha$ so that $a \in \mathcal{A}_\alpha$).

(2) Let a be the current considered element of $\mathcal{A}$. A set of feature vectors $\{\mathbf{x_i}\}$ is computed by feature extraction applied to all the regions of $\{R_i\}$. Feature extraction result depends on the features associated with a. For example, if $a = Hue$, a color coherence vector is computed for each R_i. Feature extraction result is a set of couples $\{(R_i, \mathbf{x_i})\}$ where $\mathbf{x_i}$ is the feature vector extracted from R_i.

(3) The clustering algorithm (e.g. k-means) is applied on $\{\mathbf{x_i}\}$. The result of clustering is a set of triples $\{(R_i, \mathbf{x_i}, k_i)\}$. k_i is the numerical label associated with $\mathbf{x_i}$ and R_i. $k_i = k_j$ implies that $\mathbf{x_i}$ and $\mathbf{x_j}$ belong to the same cluster.

(4) The cluster visualization and labeling step allows the user to assign a semantics to the resulting clusters. The k^{th} resulting cluster is visualized by displaying the subset of $\{R_i\}$ labeled by k. The output of cluster visualization and labeling is a training set $X = \{(\mathbf{x_i}, C_i), C_i \preceq_\Theta a \text{ and } C_i \neq a\}$. Note that modifiers (e.g. Not, Slight, Strong) provided by the visual concept ontology can be associated with visual concepts to obtain new semantic labels. The modifier Not is particularly useful to obtain negative samples of a visual concept. The resulting training set is composed of feature vectors semantically labeled by visual concepts (e.g. *Granulated*, *Smooth*, $Not(Blue)$).

During this interactive process, impure clusters may be obtained. By an impure cluster we mean that this cluster results from regions representative of several visual concepts. In this case, the clustering algorithm can be reapplied on this cluster in order to improve its purity. For instance, a cluster containing both *Smooth* and *Granulated* regions has to be splitted in two subsets in order to obtain representative samples of these visual concepts. Cluster purity is currently evaluated visually by the end-user. This approach does not require any manual segmentation and allows to label several regions at the same time.

Visual concept learning is fully automatic and consists of training a set of detectors $D = \{d_{C_i}\}$ to recognize visual concepts involved in the labeling phase. For a feature vector $\mathbf{x}$, $d_{C_i}(\mathbf{x})$ measures the confidence degree given to the hypothesis "$\mathbf{x}$ is a representative sample of C_i". Visual concept detection is seen as a two class decision problem (a one-versus-rest scheme).

Visual concept learning is composed of two steps : feature selection and training. Feature selection chooses the most characterizing features for better visual

concept detection. We use a Linear Discriminant Analysis (LDA) to perform feature selection. A support vector machine (SVM) is then trained to obtain each d_{C_i} by using the training set $X = \{(\mathbf{x_i}, C_i)\}$. To achieve training, both positive and negative samples are required. The set of positive samples of C_i is defined as the set of feature vectors labeled by $C_k \preceq_\Theta C_i$. The set of negative samples of C_i is defined as the union of the positive samples of the brothers of C_i and of the feature vectors labeled by $Not(C_i)$ during cluster labeling phase. The next section shows how $D = \{d_{C_i}\}$ is used to perform semantic indexing. The combination of the domain knowledge base (fig. 1) and visual concept detectors is called an augmented knowledge base.

5 Semantic Indexing and Retrieval

An overview of the indexing process is given in fig. 3. Semantic indexing uses a categorization algorithm divided into four steps (fig. 4).

(1) The categorization process is initiated by a **categorization request** which contains an image to index. The list of domain classes used in the algorithm corresponds to different hypotheses that have to be verified in the image.

(2) The hypothetic object of interest has to be **segmented** from the background. To achieve object extraction, we use a meanshift segmentation algorithm [11]. If the algorithm tries to classify a subpart, the segmentation task consists of extracting the subpart from the main object.

(3) Then comes **local matching** between current class attribute values (e.g. *CircularSurface* for attribute *Geometry*) and visual concepts recognized by the detectors trained during the learning process. Local matching value associated with an attribute a of a class α is defined as $m_\Theta(a) = max\{d_{C_i}(\mathbf{x})\}$ with $C_i \in \mathcal{V}_\alpha(a)$ and $m_\Theta(a) \in [0, 1]$. Feature vector $\mathbf{x}$ used to compute local matching is the result of **feature extraction**. The result of local matching is a set of confidence values associated with each attribute. For a subpart attribute, a recursive call has to be made so as to compute its global matching value.

(4) **Global matching** consists of evaluating if current class matches the object to be recognized. This matching is done by combining the results of local matching. Global matching value associated with a class α is defined as $m_\Phi(\alpha) = \sum_{a \in \mathcal{A}_\alpha} m_\Theta(a)/Card(\mathcal{A}_\alpha) + \sum_{\beta \in \mathcal{S}(\alpha)} m_\Phi(\beta)/Card(\mathcal{S}(\alpha))$. If m_Φ is

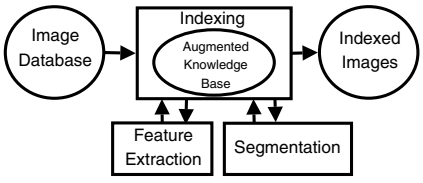

Fig. 3. The input of the indexing process is a set of images to index. The output is the same set of images coupled with semantic annotations

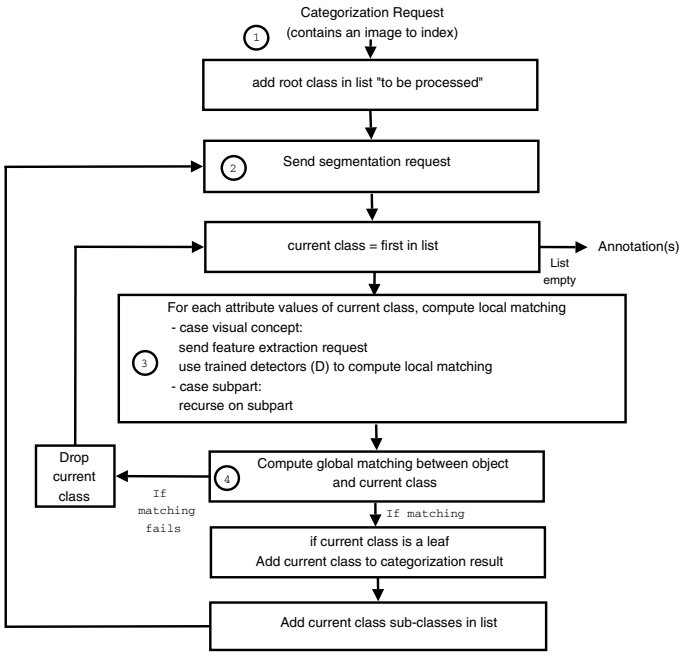

Fig. 4. Simplified version of the object categorization algorithm

greater than a predefined threshold $th_{compatibility} \in [0, 1]$ then matching between current class and unknown object is validated. If object matches current class, the classification algorithm tries to go deeper in the domain class hierarchy defined by the partial order $\preceq_{\Phi}$. If matching fails, current class is dropped.

The algorithm illustrated in fig. 4 is applied to all the images of the set of the images to index. Each image is annotated by a set of annotations (one annotation per object recognized in the image). For example, if an image of an outdoor scene is composed of sky and one aircraft, three annotations are associated with the image : one annotation for the object of class $OutdoorScene$, one annotation for the object of class Sky and one annotation for the object of class $Aircraft$.

An annotation matches the structure of a domain class and contains the following elements : the class α of the object o (e.g. $\alpha = Sky$); the mask resulting from automatic segmentation which locates o in the image; the visual description of o (i.e. the value assigned to each $a \in \mathcal{A}_{\alpha}$ associated with a confidence value) (e.g. $(Pattern = Smooth, 0.9)$); the object of which o is a subpart; the objects in spatial relation with o (e.g. if o is of class $OutdoorScene$, an object of class Sky which is related to o by the spatial relation Top).

Retrieval is initiated by a symbolic query. The output of retrieval is the subset of the indexed images which associated annotation(s) matches the query. A query is structured as a logical composition (by using the logical operators {or,and}) of the elements composing annotations. For example the query "$Class = Sky\ and\ Hue = Blue$" retrieves the images annotated as containing

blue sky. The query "$Class = OutdoorScene\ and\ Top = Sky\ and\ Bottom = Asphalt$" retrieves the images annotated as containing sky in the top part of the image and asphalt in the bottom part of the image.

6 Results

We have used an image database freely available online[1] to apply our methodology. More precisely, the following object categories have been used for learning and evaluation: motorbikes, airplanes and cars (fig. 5). Background images (fig. 5) have been used to evaluate the precision of the system. A background image is defined as not containing any object of interest. The training set is structured as following : 400 aircraft images, 200 motorbike images, 250 car images and 400 background images. The test set is structured as following : 500 aircraft images, 500 motorbike images, 250 car images and 600 background images (1850 images). No image used for training is contained in the test set.

The weakly supervised approach described in section 4 has allowed us to obtain clusters of positive and negative samples of the following visual concepts : $Blue, Grey, AircraftShape, MotorBikeShape, CarShape\ and\ Smooth$. All the images of the training set have been segmented into regions. Feature extraction, clustering and labeling have been performed for the following visual concepts of $\mathcal{A}$: $Hue, Geometry\ and\ Pattern$. The number of clusters computed by the clustering algorithm is initially set to 15. The final number of clusters may be different because of cluster purification. We have obtained about 2000 sample regions from the training set (1000 positive region samples and 1000 negative region samples) used for training the detector of the visual concept Blue (for $a = Hue$). In this case, the initial number of regions resulting from segmentation was about 11000.

A Recall/Precision curve has been obtained (by a variation of $th_{compatibility}$ from 0 to 1 with a variation step of 0.01) for the following domain classes : $Aircraft$, $MotorBike$, Car and Sky (fig. 6). Precision is defined as the ratio be-

Fig. 5. Typical images of interest on the first row: aircrafts, cars and motorbikes in their environment. Background images on the second row

[1] http://www.vision.caltech.edu/feifeili/Datasets.htm

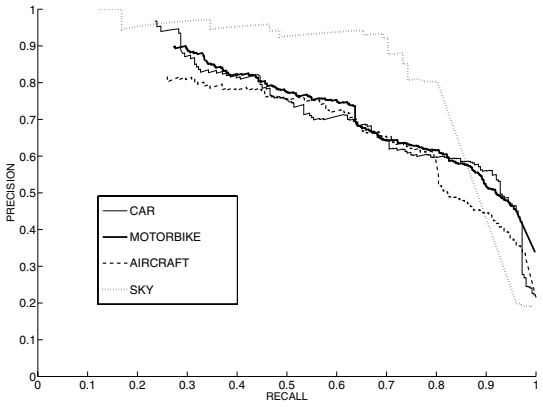

Fig. 6. Recall/Precision curves obtained for some domain classes

tween the number of relevant retrieved images and the number of retrieved images. Recall is defined as the ratio between the number of relevant retrieved images and the number of relevant images in the image database. The results obtained show that our methodology leads to efficient indexing : For a recall of 0.5, precision is between 0.75 and 0.78 for the domain classes $Aircraft$, $MotorBike$ and Car and of 0.90 for class Sky. These results show that even with very little effort of knowledge acquisition (6 visual concepts and 4 domain classes), the approach offers both good results and semantic richness.

7 Conclusion and Future Works

We have presented a new approach for semantic image indexing and retrieval. Our approach is based on both knowledge based techniques and machine learning techniques. A priori knowledge is structured as a hierarchy of domain classes described by visual concepts provided by a visual concept ontology. This ontology provides an easy access to a wide range of low-level image processing algorithms (e.g. color, texture and shape analysis algorithms). From a set of image samples, a weakly supervised learning phase allows to obtain region samples of the visual concepts used during knowledge acquisition. These region samples are used to train visual concept detectors capable of visual concept detection in any image. Semantic indexing uses these visual concept detectors to produce symbolic annotations of the images to index. During the indexing phases, the visual concepts allow the system to extract the most distinctive visual characteristics for better recognition of the domain classes. We have shown that our approach leads to efficient image indexing. The semantic nature of the annotations enables the user to express queries at a conceptual level that is difficult to reach with classic query-by-example paradigm. Moreover, the retrieval process does

not have to cope with the issues (i.e. scalability and performance) encountered with numerical databases.

In the short term, we aim at improving the weakly-supervised phase by using hierarchical clustering techniques which should ease cluster labeling. We also aim at improving the retrieval phase by making better use of a priori knowledge. Another important remaining challenge is to achieve semantically driven segmentation which would use the visual description of the domain classes to choose the most adapted segmentation algorithms and to improve the splitting/merging of the image data resulting from segmentation.

References

1. Town, C., Sinclair, D.: Language-based querying of image collections on the basis of an extensible ontology. IVC **22** (2004) 251–267
2. Soo, V.W., Lee, C.Y., Li, C.C., Chen, S.L., Chen, C.C.: Automated semantic annotation and retrieval based on sharable ontology and case-based learning techniques. In: JCDL '03, IEEE Computer Society (2003) 61–72
3. Fauqueur, J., Boujemaa, N.: New image retrieval paradigm: logical composition of region categories. In: ICIP03. (2003) III: 601–604
4. Li, J., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. Pattern Anal. Mach. Intell. **25** (2003) 1075–1088
5. Mezaris, V., Kompatsiaris, I., , Strintzis, M.: Region-based image retrieval using an object ontology and relevance feedback. EURASIP JASP **2004** (2004) 886–901
6. Maillot, N., Thonnat, M., Boucher, A.: Towards ontology based cognitive vision. Machine Vision and Applications (MVA) **16** (2004) 33–40
7. Pass, G., Zabih, R., Miller, J.: Comparing images using color coherence vectors. In: ACM Multimedia. (1996) 65–73
8. Zhang, J., Tan, T.: Brief review of invariant texture analysis methods. Pattern Recognition **35** (2002) 735–747
9. Csurka, G., Dance, C., Bray, C., Fan, L., Willamowski, J.: Visual categorization with bags of keypoints. In: Pattern Recognition and Machine Learning in Computer Vision Workshop, Grenoble, France (2004)
10. Boder, M.: Mpeg-7 visual shape descriptors. IEEE Transactions on Circuits and Systems For Video Technology **11** (2001) 716–719
11. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. PAMI **24** (2002) 603–619