Multi-stage Classification of Images from Features and Related Text

John R. Smith
IBM T.J. Watson Research Center
30 Saw Mill River Road
Hawthorne, NY 10532
jrsmith@watson.ibm.com

Shih-Fu Chang
Dept. of Electrical Engineering
Columbia University
New York, NY 10027
sfchang@ctr.columbia.edu

Abstract

The synergy of textual and visual information in Web documents provides great opportunity for improving the image indexing and searching capabilities of Web image search engines. We explore a new approach for automatically classifying images using image features and related text. In particular, we define a multi-stage classification system which progressively restricts the perceived class of each image through applications of increasingly specialized classifiers. Furthermore, we exploit the related textual information in a novel process that automatically constructs the training data for the image classifiers. We demonstrate initial results on classifying photographs and graphics from the Web.

1 Introduction

The tremendous proliferation of visual information in the World-Wide Web is increasing the need for more sophisticated methods for automatically analyzing, interpreting and cataloging this imagery. The recent development of content-based query systems has advanced our capabilities for searching for images by color, texture and shape features [FSN+95, BFG+96, SC96]. However, these systems are limited in their capability for automatically assigning meaningful semantic labels to the images.

In this paper, we present a method for classifying images using image features and related textual information. We focus on the World-Wide Web, where a large variety of imagery consisting of graphics, animations, photographs, and so forth, is published in Web documents. The multi-stage classification system provides a hierarchy of classifiers that are trained from the images on the Web that are sufficiently annotated by text. In the successive stages, the classes are restricted as the classifiers utilize more complex features and increased training.

1.1 Related work

The classification of images in the World-Wide Web has been explored in [RF97, ASF97, FMF+96, SC97]. In [ASF97], multiple decision trees based upon image feature metrics are used for distinguishing photographs and graphics on the Web. The results are used to enhance the image search capabilities of the Webseer system. Alternatively, in order to better index the images in Web documents, Rowe and Few are developing a system for automatically associating the text in the Web documents with the corresponding images [RF97]. In [FMF+96], the images are analyzed using a blob-world representation in which objects such as people and animals are detected by matching the blobs to pre-defined body plan templates. In [SC97], as part of the WebSEEk image and video search engine, we developed a system for classifying images into subject classes using text derived from image addresses and HTML tags. We now extend this classification system to utilize image features.

1.2 Multi-stage classification system

The multi-stage image classification system consists of three stages as illustrated in Figure 1. Each stage utilizes image features and/or text. In the first stage, the images are classified into type classes, i.e., color photos, graphics, gray photos, using a decision tree based upon the analysis of image features in HSV color space. In the second stage, the images are further classified into more restricted composition classes, i.e., silhouettes, center-surround images, scenes, and textures using more complex features derived from image spatial sampling and region extraction. Finally, in the last stage, the images are classified into

semantic classes, i.e., beaches, buildings, nature, sunsets, and so forth, using specialized classifiers which are trained from images that are classified from their related text.

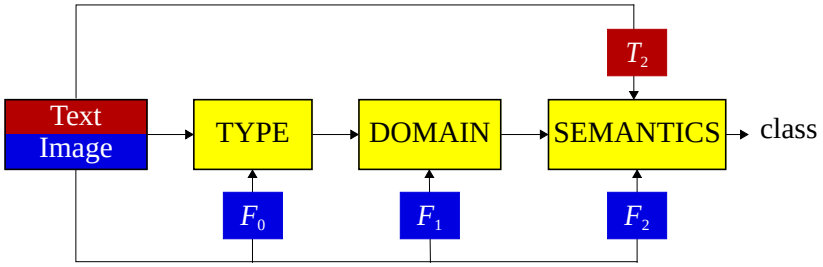

Figure 1: Multistage image classification system uses image feature sets: $\mathcal{F}_0$, $\mathcal{F}_1$, $\mathcal{F}_2$, and related text $\mathcal{T}_2$.

In this paper, we present the multi-stage image classification system and describe the processes for classifying the images into the type, composition and semantic classes. In Section 2, we introduce a new simple feature decision tree for determining image type. We present, in Section 3, the image composition classification system. Finally, in Section 4, we present a novel semantics classification system, which uses composite region templates (CRTs). We evaluate the performance of the CRT-based semantics classification system in classifying images from eight semantics classes.

2 Stage 1 – image type

In the first stage, images are classified into image type classes. The image type hierarchy is illustrated in Figure 2. We define the following set of image type classes: color photographs, complex color graphics, simple color graphics, gray photos, gray graphics, b/w (bi-level) photographs, and b/w graphics. The type classes are given by the root nodes of the decision tree in Figure 2.

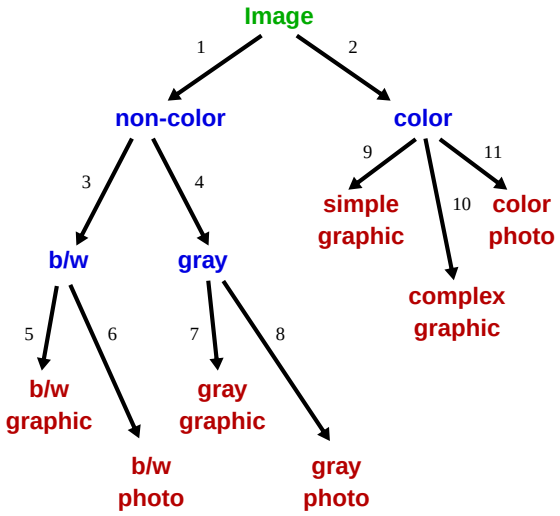

Figure 2: Image type hierarchy with five decision points: $(1, 2)$, $(3, 4)$, $(5, 6)$, $(7, 8)$, $(9, 10, 11)$.

2.1 Image type features

In order to automatically classify the images into type classes, the system analyzes the image features in HSV color space. The transformation and quantization to 166 HSV colors is given in [Smi97]. The following HSV color features are extracted from the images:

- A = relative amount of black,
- B = relative amount of white,

- C = relative amount of gray,
- D = relative amount of colors which are fully saturated, i.e., saturation = 1,
- E = relative amount of colors which are half saturated, i.e., saturation ≥ 0.5 and saturation < 1,
- F = number of colors present from the 166-color quantized HSV color space,
- G = number of grays present from the 166-color quantized HSV color space,
- H = number of hues present from the 166-color quantized HSV color space,
- I = number saturations present from the 166-color quantized HSV color space.

Table 1 gives the average feature values for the image type classes obtained from a training set of several thousand images retrieved from the World-Wide Web.

Image type	A black	B white	C gray	D fully sat.	E half sat.	F # colors	G # grays	H # hues	I # sats.
color photo	0.18	0.06	0.14	0.04	0.10	11.9	111	54	94
complex graphic	0.07	0.03	0.06	0.18	0.23	29.8	77	76	80
simple graphic	0.16	0.26	0.18	0.16	0.07	3.8	17.8	8.0	14.4
gray photo	0.24	0.06	0.70	0	0	0	130	1	1
gray graphic	0.21	0.29	0.49	0	0	0	23.2	1	1
b/w photo	0.60	0.40	0	0	0	0	2	1	1
b/w graphic	0.41	0.59	0	0	0	0	2	1	1

Table 1: Image type classes and corresponding attributes obtained from training images.

Starting at the root node in the decision tree, images are classified into increasingly specific type classes. In order to derive the decision criteria, we computed the image color features for a large set of training images. For each decision point, we identified the subset of features that were relevant to that decision. For example, for decision point $(1, 2)$, image features A, B and C are sufficient.

For each decision point, a multi-dimensional space was generated, such that each dimension corresponds to one of the relevant features (i.e., A, B, C). This multi-dimensional space was then partitioned adaptively to the training images. The frequencies by which training images of each type occur within the partitions determines the decision criteria. In this way, a new image is quickly classified by simply obtaining the most likely class in the partition corresponding to its feature values.

2.2 Adaptive partitioning

The M dimensional decision space is iteratively partitioned as follows, where τ is a training threshold ($\tau = 0.9$):

1. Assign training images to points in the M dimensional decision space by measuring their feature values.
2. Assign initial partition R_0 to the entire M dimensional decision space.
3. Split R_0 into 2^M partitions by bi-secting R_0 along each dimension.
4. For each new partition R_j, if $\neg\exists C_k$ such that $P(C_k|R_j) > \tau$ then split R_j, and repeat Step 3 and 4 as necessary.
5. For each partition R_l after all splitting, assign the likelihood of each class C_k to each partition R_l as follows:

$$P(C_k|R_l) = \frac{P(R_l|C_k)P(C_k)}{P(R_l)},$$

where $P(R_l|C_k)$ is the number of training points in partition R_l that belong to class C_k, $P(C_k)$ is the number of training points from class C_k, and $P(R_l)$ is the number of points in partition R_l.

2.3 Type classification

Given the partitioned decision space, the type class of an unknown image is determined by simply looking up which class C_k maximizes $P(C_k|R_l)$, where R_l is the partition corresponding to the features of the unknown image.

3 Stage 2 – image composition

In the second stage, the images are assigned to one of the following composition classes: silhouettes, center-surround images, scenes and textures. The image composition is determined by the separation of the center and surround areas in the image.

3.1 Center-surround separation

The image center and surround are separated by using two methods of sampling the surround areas of the image, depicted in Figure 3 as regions 'A,' 'B,' 'C,' and 'D.'

1. Method 1: most prominent color – From regions A, B, C, D, the most prominent color in the surround, i.e., given by m, where, $\forall m \neq k, h_S[m] \geq h_S[k]$, is back-projected onto the image (see [Smi97] for details about back-projection) to extract the surround region, depicted in Figure 3 as S_1.

2. Method 2: pooled color histogram – From regions A, B, C, D, a pooled color histogram is generated as follows: $\mathbf{h}_S = \mathbf{h}_A + \mathbf{h}_B + \mathbf{h}_C + \mathbf{h}_D$. Then $\mathbf{h}_S$ is back-projected onto the image to more completely extract the surround region, depicted in Figure 3 as S_2.

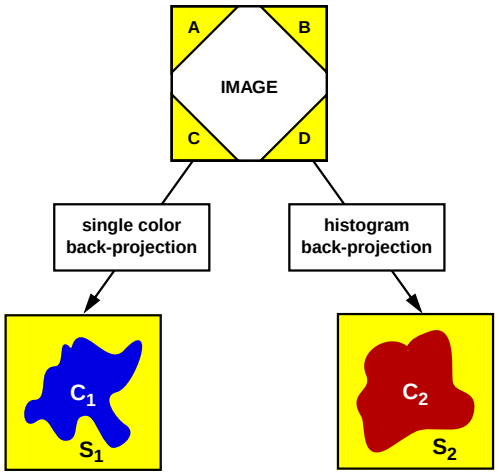

Figure 3: Image center-surround separation process for image composition classification extracts two versions of the center regions (C_1, C_2) and surround regions (S_1, S_2).

Method 1 (back-projecting the most prominent surround color) is more suited for extracting a silhouetted object that is depicted on a single color background. Method 2 (back-projecting the pooled surround histogram $\mathbf{h}_S$) is more suited for separating a multi-color surround from a center region. The results of the back-projections yield two versions of a center object, denoted by C_1 and C_2. The attributes of the extracted center regions (C_1, C_2) and surround regions (S_1, S_2) are used to determine the image composition class.

The attributes used for image composition classification are derived from the sizes of C_1 and C_2, and the color distances between C_1 and S_1, and C_2 and S_2, respectively. Table 2 indicates the typical values of the image features used for composition classification. The 'size' features indicate the relative sizes of the extracted image center regions. The 'dist' features indicate the distances in HSV color space between the respective center and surround regions.

Figure 4 illustrates the results from the center-surround separation process for the four image composition classes. For the silhouette images, Methods 1 and 2 produce similar results since the surround

Image composition	size(C_1)	size(C_2)	dist(C_1, S_1)	dist(C_2, S_2)
silhouette	0.59	0.58	0.89	0.68
center-surround	0.54	0.23	0.69	0.54
scene	0.83	0.19	0.40	0.23
texture	0.14	0.05	0.19	0.12

Table 2: Image composition classes and corresponding center-surround features.

typically contains a single color. For the center-surround images, Method 2 extracts a larger surround than Method 1 since the surround contains more than one color. Furthermore, the color distance between the center and surrounds in both cases is relatively large. In the case of the scene images, Method 2 extracts a large surround region while method 1 extracts a small surround region. Finally, for textures, both methods fail at separating a center from the surround.

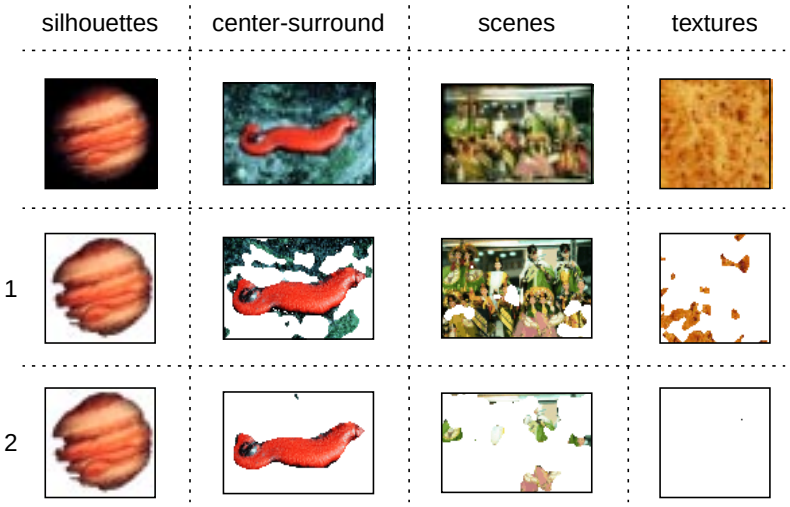

Figure 4: Center-surround separation examples using Methods 1 and 2 for the four image composition classes.

3.2 Composition classification

Given the image composition feature set, the decision space is derived from training images using adaptive partitioning of the 4-dimensional feature space. The classification of an unknown image is performed by simply extracting the center-surround features and finding the partition corresponding to the feature values. Similar to the case for image type classification, the composition label is assigned by the most likely composition class in the partition.

4 Stage 3 – image semantics

In the final stage, the images are classified into semantics classes derived from a semantics ontology (described in [SC97]). Here, we examine eight semantics classes: beaches, buildings, crabs, divers, horses, nature, sunsets, and tigers.

4.1 Text-to-subject mapping

The semantics classes are defined by identifying training images on the Web that are associated with relevant text. These images are assigned to the semantics classes by mapping the key-terms to semantics classes[1]. For example, the key-term 'sunset' is mapped into semantics class 'nature/sunsets.' This process

[1] The WebSEEk demo: `http://disney.ctr.columbia.edu/webseek`

is described in more detail in [SC97]. We now describe how the images that cannot be semantically classified using text due to lack of useful annotations, are classified using images features based upon composite region templates.

4.2 Composite region templates

The composite region templates (CRTs) are defined from training images from the semantic classes. The system extracts the color regions from the images and generates a set of region strings for each semantic class. The region strings for each class are then consolidated into the sets of CRTs.

4.2.1 Region string generation

The region strings are generated in a series of five vertical scans of the image which order the extracted regions from top-to-bottom. The five vertical scans are equally spaced horizontally. Since the images are normalized to 100 pixels, each vertical scan covers a 20-pixel wide area. In each scan, the symbol value of each consecutive region is concatenated onto the scan's region string. In general, the symbol values (i.e., symbol 'A,' 'B,' 'C' in Figure 5) represent the index values of the features of the regions.

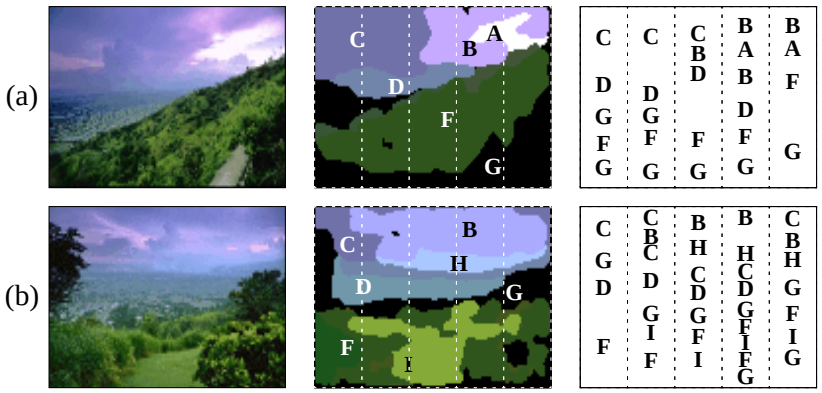

Figure 5: Examples of region extraction and region string generation using a top-to-bottom orderings (a) (CDGFG, CDGFG, CBDFG, BABDFG, BAFG), (b) (CGDF, CBCDGIF, BHCDGFI, BHCDGFIFG, CBHGFIG).

An example of the region string generation process for two nature images is illustrated in Figure 5. We can see that for the two nature images, the symbols 'A,' 'B,' and 'C' (sky) typically preceed symbols 'F,' and 'G' (grass). The objective of the CRT method is to detect these important relationships between regions for each semantic class. The top-to-bottom scans capture the relative vertical placement of the regions. Note that the five region strings from an image are not subsequently distinguished by the horizontal position of the scan.

Definition 1 Region String. *A region string* $\mathbf{S}$ *is a series of symbols* $\mathbf{S} = s_0 s_1 s_2 \ldots s_{N-1}$*, which is generated from the regions of an image where* s_n *is the symbol value (i.e., color index value) of the* n^{th} *successive region in a top-to-bottom scan.*

4.2.2 Region string consolidation

After the region strings are generated, they are consolidated to generate the CRTs in order to capture the recurring arrangements of the regions within the images and semantic classes. The CRTs characterize, in general, the order of the symbols in the region strings but not their adjacency. The likelihood of these CRTs within and across the semantics classes forms the basis of the semantics classification system.

Definition 2 CRT. *A composite region template* $\mathbf{T}$ *is an ordering of* M *symbols,* $\mathbf{T} = t_0 t_1 t_2 \ldots t_{M-1}$.

The region strings are consolidated by detecting and counting the frequencies of the CRTs in the set of region strings. For example, the test for $\mathbf{T} = t_0 t_1 t_2$ in region string $\mathbf{S}$ is given by $I(\mathbf{T}, \mathbf{S})$, where

$$I(\mathbf{T}, \mathbf{S}) = \begin{cases} 1 & \text{if } s_l = t_0 \text{ and } s_m = t_1 \\ & \text{and } s_n = t_2 \text{ and } l \leq m \leq n \\ 0 & \text{otherwise.} \end{cases}$$

The frequency of each CRT, $\mathbf{T}_i$, in a set of region strings $\{\mathbf{S}_j\}$ is then given by $P(\mathbf{T}_i)$, where

$$P(\mathbf{T}_i) = \sum_j I(\mathbf{T}_i, \mathbf{S}_j).$$

The frequency of each CRT, $\mathbf{T}_i$, in the set of region strings $\{\mathbf{S}_j\}_k$ from semantic class C_k is given by $P(\mathbf{T}_i|C_k)$, where

$$P(\mathbf{T}_i|C_k) = \sum_{\forall_j \mathbf{S}_j \in C_k} I(\mathbf{T}_i, \mathbf{S}_j).$$

4.2.3 CRT library

The CRTs derived from the training images construct the CRT library, which is defined as follows:

Definition 3 CRT library. *A composite region template library is given by a set of* $(K+2)$*-tuples:*

$$\{\mathbf{T}_i, P(\mathbf{T}_i), P(\mathbf{T}_i|C_0), P(\mathbf{T}_i|C_1), \ldots, P(\mathbf{T}_i|C_{K-1})\},$$

where K *is the number of semantic classes.*

4.3 Decoding image semantics

Once the CRT library is built from training images, it is used to semantically classify the unknown images. The semantics of an unknown image are decoded from its set of region strings using the CRT library as follows:

1. First, the region strings for the unknown image are extracted and consolidated into a set of CRTs.
2. For each CRT, $\mathbf{T}'_i$, from the unknown image, $P(C_k|\mathbf{T}'_i)$ is computed from the entries in the CRT library from:
$$P(C_k|\mathbf{T}'_i) = \frac{P(\mathbf{T}'_i|C_k)}{P(\mathbf{T}'_i)} P(C_k).$$
3. The classification of the unknown image is then given by: assign image to class l when
$$\forall_{l \neq k}, \quad \sum_i P(C_l|\mathbf{T}'_i) > \sum_i P(C_k|\mathbf{T}'_i). \tag{1}$$
That is, class C_l best explains the CRTs represented in the region strings of the unknown image.

4.4 Semantics classification evaluation

We evaluate the CRT-based semantics decoding method by measuring its performance in classifying unknown images from the eight semantic classes. Example images are illustrated in Figure 6. In the experiments, images from eight semantic classes were classified using the CRT method.

In total, 261 images were identified as belonging to the eight semantic classes. These 261 images were divided into non-overlapping training and test sets according to Table 3. The system used the 71 training images to generate the CRT library. The remaining 190 test images were used to evaluate the semantics classification performance of the system. The classification results are given in Table 3.

Given the eight semantics classes, the semantics decoding system using CRTs provides a classification rate of 0.784. The majority of classification errors resulted from a confusion between the buildings and nature classes. This is not surprising since both classes, as illustrated in Figure 6, often depict similar scenes, such as blue skies, above brown objects, above green grass.

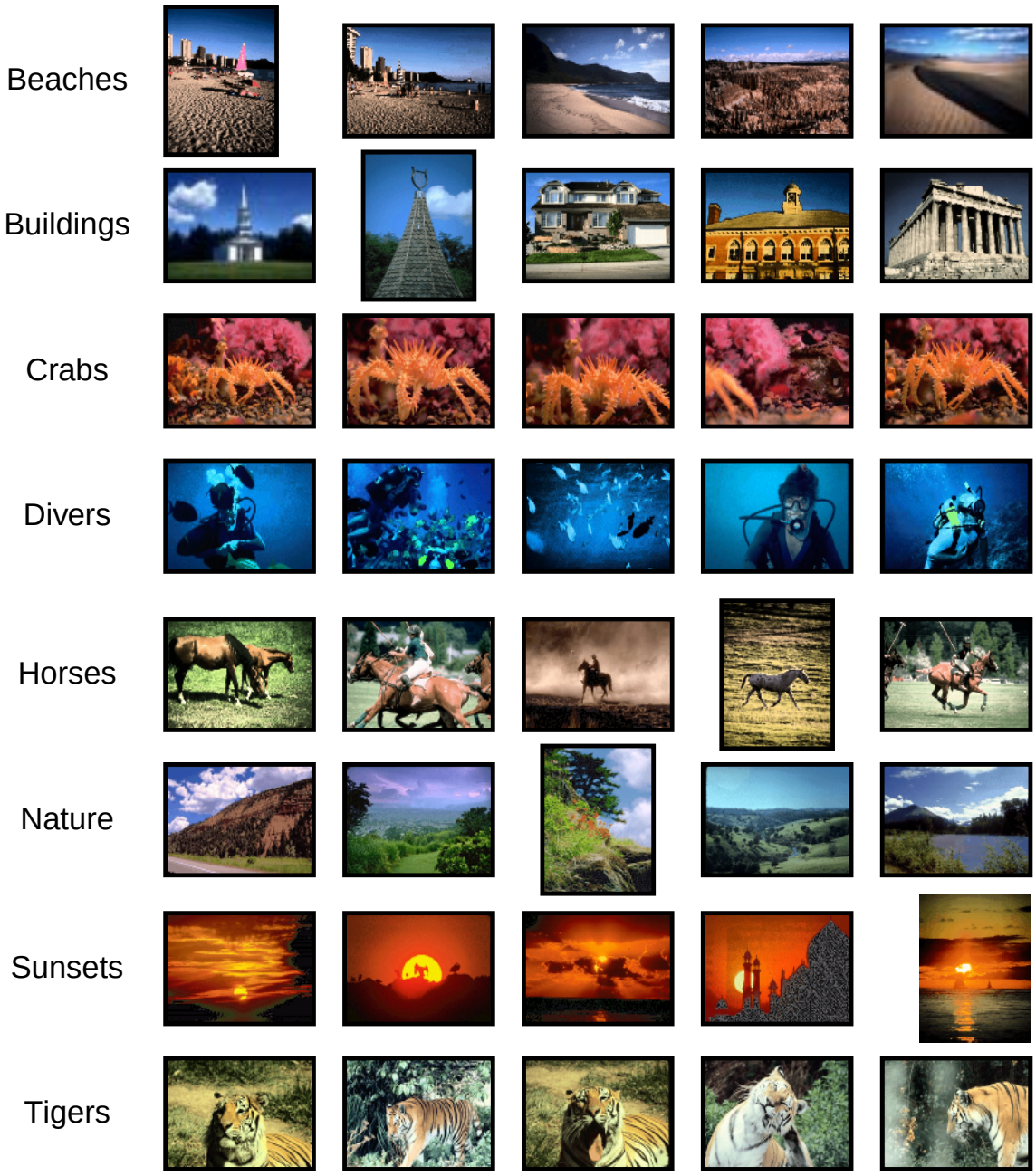

Figure 6: Example images from the eight semantics classes used to evaluate the CRT semantics decoding system: beaches, buildings, crabs, divers, horses, nature, sunsets, and tigers.

	overall	beaches	buildings	crabs	divers	horses	nature	sunsets	tigers
# total	261	14	56	9	33	26	46	46	31
# train	71	7	10	4	10	10	10	10	10
# test	190	7	46	5	23	16	36	36	21
# correct	149	6	30	5	23	14	20	31	21
% correct	78.4	85.7	65.2	100	100	87.5	55.6	86.1	100

Table 3: Image semantics classification experiment results using 71 training images and 190 test images from eight semantics classes.

5 Summary and Future Work

We presented a new system for classifying images using features and related text. The multi-stage image classification assigns the images to type, composition and semantics classes. Image type and composition are determined by mapping image features into a decision space that is adaptively partitioned using training images. Image semantics are determined by a novel system which matches the arrangements of regions in the images to composite region templates (CRTs). We developed a process by which this CRT library is constructed automatically from the images that are textually annotated.

We are applying the multi-stage image classification system to the classification of images on the World-Wide Web in order to better index and catalog this visual information. In particular, we are investigating the performance of the image semantics decoding system using several thousand semantics classes. Finally, we are exploring the utility of the image classification system for customizing the delivery of Web documents.

References

[ASF97] V. Athitsos, M. J. Swain, and C. Frankel. Distinguishing photographs and graphics on the World Wide Web. In *Proceedings, IEEE Workshop on Content-based Access of Image and Video Libraries*, June 1997.

[BFG+96] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. C. Jain, and C. Shu. Virage image search engine: an open framework for image management. In *Symposium on Electronic Imaging: Science and Technology – Storage & Retrieval for Image and Video Databases IV*, volume 2670, pages 76 – 87. IS&T/SPIE, January 1996.

[FMF+96] D. A. Forsyth, J. Malik, M. M. Fleck, T. Leung, C. Bregler, C. Carson, and H. Greenspan. Finding pictures of objects in large collections of images. In *Proceedings, International Workshop on Object Recognition*. IS&T/SPIE, April 1996.

[FSN+95] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The QBIC system. *IEEE Computer*, 28(9):23 – 32, September 1995.

[RF97] N. C. Rowe and B. Frew. Automatic caption localization for photographs on World Wide Web pages. Technical Report Code CS/Rp, Dept. of Computer Science, Naval Postgraduate School, 1997.

[SC96] J. R. Smith and S.-F. Chang. VisualSEEk: a fully automated content-based image query system. In *Proc. ACM Intern. Conf. Multimedia*, pages 87 – 98, Boston, MA, November 1996. ACM.

[SC97] J. R. Smith and S.-F. Chang. Visually searching the Web for content. *IEEE Multimedia*, 4(3):12 – 20, July–September 1997.

[Smi97] J. R. Smith. *Integrated Spatial and Feature Image Systems: Retrieval, Analysis and Compression*. PhD thesis, Graduate School of Arts and Sciences, Columbia University, New York, NY, 1997.