Multi-Level Visual Alphabets

Menno Israël[1,*], Jetske van der Schaar[2,*,**], Egon L. van den Broek[3,**], Marten den Uyl[4] and Peter van der Putten[5]
[1] Netherlands Forensic Institute, P.O. Box 24044, 2490 AA The Hague, The Netherlands
e-mail: menno@nfi.nl
[2] Radio Netherlands Worldwide, Witte Kruislaan 55, 1217 AM Hilversum, The Netherlands
e-mail: jetske.vanderschaar@rnw.nl
[3] Human Media Interaction, Faculty of EEMCS, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: vandenbroek@acm.org
[4] VicarVision, Singel 160, 1015 AH Amsterdam, The Netherlands
e-mail: denuyl@vicarvision.nl
[5] LIACS, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands
e-mail: putten@liacs.nl

***Abstract*— A central debate in visual perception theory is the argument for indirect versus direct perception; i.e., the use of intermediate, abstract, and hierarchical representations versus direct semantic interpretation of images through interaction with the outside world. We present a content-based representation that combines both approaches. The previously developed Visual Alphabet method is extended with a hierarchy of representations, each level feeding into the next one, but based on features that are not abstract but directly relevant to the task at hand. Explorative benchmark experiments are carried out on face images to investigate and explain the impact of the key parameters such as pattern size, number of prototypes, and distance measures used. Results show that adding an additional middle layer improves results, by encoding the spatial co-occurrence of lower-level pattern prototypes.**

***Keywords*— Visual perception, visual alphabets, content based image retrieval.**

I. Introduction

A broad range of attempts have been made to bridge the semantic gap in multi-media retrieval. Recently, [4] described this challenge as follows:

. . . narrowing the large disparity between the low-level descriptors that can be computed automatically from multi-media content and the richness and subjectivity of semantics in user queries and human interpretations of audiovisual media - the so-called Semantic Gap. (p. 137)

This statement did not just address the gap between syntactic and semantics – it also hints at the contrast between low level features and rich concepts. Should there be a layering of representations in between?

Humans and animals process the visual world by combining a rich set of cues such as size, shape, color, lightness, motion, depth into a useful representation of the natural environment [17]. Early models from both computer and human vision propose a sequence of layers of representation, consisting of abstract geometrical features of increasing complexity [13]. In contrast, proponents of direct perception argued that geometrical features and abstract representations are unnecessarily complex and disconnected from the environmental niche an animal lives in [7].

The debate of direct versus indirect perception is still relevant today [1], [20]. Even if the purpose is to develop multi-media applications rather than model cognitive systems, it can be useful to refer to cognitive theories of perception for inspiration. In this article, we present an approach to semantic image representation and classification that aims to combine aspects of both views on perception. We introduce a hierarchy of representations, each level feeding into the next one, with features derived from and tuned to the environment itself. Note whilst our method is biologically inspired we claim no relevance whatsoever to understanding human or animal visual perception.

The method that will be presented is an extension of our earlier approach to scene classification: Visual Alphabets [8], [9]. End users (i.e., domain experts, not image processing experts) collect examples of image fragment classes (e.g., sky, grass, or bricks) relevant for recognizing settings at the image level (e.g., countryside versus city). Subsequently, models are constructed to classify image fragments as well as the images itself. We have shown that this is a generic method for building task specific setting classifiers, without requiring specialist image processing knowledge, and described successful applications of our method ranging from television archive search and navigation, sewage inspection by robots and internet porn filtering [8], [9].

To enable an extension of the Visual Alphabets method to other domains (e.g., forensics, homeland security, and deeper media archive indexing), a drawback of the original Visual Alphabets method has to be tackled. The main limitation of the Visual Alphabets method was that it was devoted to scenes, or in terms of the conceptual distinction introduced by Picard and Minka: 'stuff' rather than 'things' [15]. This article will

* At the time this research was conducted, the author was affiliated with VicarVision, Amsterdam, The Netherlands.

** At the time this research was conducted, the author was affiliated with the Division of Cognitive Artificial Intelligence, Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands.

Fig. 1. An image and corresponding micro, meso and macro patterns.

outline an extension of the Visual Alphabets method to multi-level Visual Alphabets. By analogy, we take the step from letters to letter sequences or words to sentences.

The approach is generalizable to any number of levels, in this paper we start with a three layer approach, with the first two levels learnt in an unsupervised manner, using positive examples only. At the lowest level, groups of pixels form a micro pattern, and we extract micro pattern prototypes through clustering. Similarly, one level up images are represented by meso patterns: areas consisting of patterns of micro pattern prototypes, extracted through clustering. Finally, macro patterns are constructed from the meso pattern representation, which can be used for classification. This is modeled as a supervised item set mining task, for which we have applied a stochastic approach combining simulated annealing for search with the diverse density multi instance learning measure for evaluation [11]. Figure 1 outlines multi level Visual Alphabets applied to face recognition, with meso pattern prototypes representing parts of a nose, eye, and hair.

Our original Visual Alphabet representation was mainly inspired by classical work in scene classification [5], [14], [15] and visual codebooks in general [10], [20], [22]; see [8], [9] for full references. The original representation differs from most related work in that it uses the local patch classifications solely as input for the classification of the scene as a whole. Hierarchical representations, using either visual alphabets or codebooks are rare. One of our main goals is to investigate whether or not useful salient pattern prototypes can evolve at intermediate levels. We also compare our method with approaches that include representations of spatial relationships across features [2], [12], [21].

The remainder of this paper consists of an overview of our method (Section II), experimental results (Section III) and a conclusion (Section IV).

II. Multi-level visual alphabet representations

In this section, we provide an overview of our method, from the micro level up to the top macro level. The approaches taken at the micro and meso level are actually quite similar. Although a single meso level is used, one could envisage adding any number of meso levels in between; that said, this is out of scope for this paper.

A. Micro Pattern Discovery

Our approach allows for using feature dimensions of choice, as this will not change the overarching framework. See for instance [16], for guidelines on selecting proper feature dimensions. In our experiments, we have primarily settled on hue (H), intensity (I), and texture dimensions. H is typically used to represent the main property of the percept color, described with labels such as yellow and red [18], [19], H is independent from I and S [19]. We derived H from RGB values as follows:

$$H = \cos^{-1}\left[\frac{[(R-G)+(R-B)]}{2\times\sqrt{(R-G)^2+(G-B)\times(R-B)}}\right] \quad (1)$$

For I, a single accepted definition exists: the average of the R, G, and B values, as proposed by [16], [19]. S is disregarded as it contains little information. For a given pattern, histograms can be calculated to reflect the distribution of H and I values. This is done using a smoothed histogram approach, as introduced in [9]. Note that the H dimension is cyclical; see also Equation 1. Texture is represented through three features: i) variance of pixels, computed over a pixel versus its neighboring values, ii) dominant direction, and iii) strength of the direction. See [9] for a formal description.

To generate the candidate set of micro patterns, we randomly sample fixed size micro patterns from a given set of images of a specific image class and calculate the corresponding feature histograms. For simplicity, this approach was preferred over other heuristics. For instance, [2] and [21] only extract patterns with sufficient levels of structure.

In the final step, we extract the relevant patterns from the candidate set for an image class, as the candidate set of micro patterns can grow very large. Depending on micro pattern size and number of images, the need for discriminating patterns has to be balanced with the objective of reducing ambiguity and overlap in the representation. We cluster the candidate patterns found with k-means clustering. Empty clusters get reinitialized with the pattern of the worst matching pattern. We experimented with a variety of distance metrics including Euclidean distance, normalized correlation:

$$1-\left|\frac{1}{n-1}\sum_n\frac{(m_{in}-\bar{m}_i)\times(m_{jn}-\bar{m}_j)}{\sigma_{m_i}\times\sigma_{m_j}}\right|, \quad (2)$$

and Mahalonobis distance:

$$\sqrt{\sum_n(m_{in}-m_{jn})S^{-1}\sum_n(m_{in}-m_{jn})}, \quad (3)$$

with a n-dimensional feature space, m_x being pattern vector x, $\bar{m}_x$ being the average pattern vector of x, σ_x as the standard deviation of x, and S^{-1} is the inverted covariance-matrix.

B. Meso Pattern Discovery

A meso pattern can be described as a pattern of micro patterns - or better as a pattern of micro pattern cluster prototypes. For meso pattern discovery, a process similar to the process for micro patterns is followed. An image is reconstructed by its micro patterns, using only the cluster centers. Next, meso patterns of a given size are sampled randomly from this micro pattern representation of the image to encode potentially meaningful spatial relationships. The feature representation is a normalized histogram of micro pattern frequencies. Subsequently, these meso patterns are clustered using k-means clustering, with the distance measure of choice (see Equations 2 and 3) to extract prototype meso patterns. In line with the micro patterns, this is only done for positive images; i.e., images that belong to the image class of interest.

C. Macro Pattern Discovery

Whilst the procedures used at micro and meso level are similar, the approach to the macro level deviates from these. The goal of the macro level is to find macro patterns that can discriminate between image classes, which can be seen as a supervised item set mining task. However, this means we cannot use standard item set mining algorithms such as APRIORI without modifications [3]. APRIORI exploits the fact that an item set can never be more frequent than any of its subsets; however, in our case the item set can be more differentiating. Various search approaches have been introduced for content-based indexing; e.g., gradient ascent [12] and greedy search [21]. For this research, however, we chose to apply a simulated annealing approach to avoid local optima.

We initialize the process by selecting a random meso pattern from a random positive class image to be a macro pattern. Then, the simulated annealing cycle starts. First, all candidate patterns are identified that differ only a single item from the current pattern. We generate all item sets (i.e., sets of meso patterns) that both differ on only one meso pattern from the current pattern and appear in any positive image. We randomly select one of these patterns. If it is better than the current best pattern, it replaces this; otherwise, it replaces it with a certain probability. This probability slowly declines to manage exploration versus exploitation. This process is repeated by selecting a new meso pattern from the positive images.

For the evaluation function we have adopted the diverse density method [12]: a method for multiple instance learning tasks. In these tasks, the goal is to learn a concept, using positive and negative bags of instances. A bag is labeled positive if at least one element is positive and negative if all elements are negative. This lends itself well to our domain as a single face meso pattern occurring can be a strong indicator for a face regardless of non face meso patterns being present.

Let B_i^+ be a positive bag and B_{ij}^+ be the j-th instance in that bag. Similarly, B_{ij}^- is an instance from a negative bag B_i^-. If the true concept were a single point t, this could be found through maximizing $P(x = t|B_1^+, \ldots, B_n^+, B_1^-, \ldots, B_m^-)$ over all x in instance space. Assuming an uninformative prior over the target and conditional independence of the bags given target t we can apply Bayes rule twofold and prove this is equivalent of finding x such that

$$argmax_x \prod_i P(x = t|B_i^+) \prod_i P(t|B_i^-). \quad (4)$$

Using the key idea that one element that is positive is sufficient for the bag to be positive we can model $P(x = t|B_i^+)$ as the probability that not all points missed the target; i.e., $P(x = t|B_i^+) = 1 - \prod_j (1 - P(x = t|B_{ij}^+))$ and equally $P(x = t|B_i^-) = 1 - \prod_j (1 - P(x = t|B^- S_{ij}))$. See [12] for full derivations and details.

III. Experiments and Results

The emphasis of the explorative experiments is on explaining the influence of the parameters on performance: pattern size, distance measure used and number of clusters, at micro, meso, and macro level. Consequently, the added value of an additional meso level is expected to be revealed. Additionally, the results were compared with those of similar methods [2], [12], [21]. This served as a minimal test to show that our method delivers at least reasonable results, rather than aiming to prove that it is superior, given that our results are based on different benchmark data.

A. Experimental Set Up

A data set of 400 images (480×360 pixels, 24 bit color depth) was used; 200 positives, consisting of Corel Gallery images (39), a UK company database with photo IDs for access control (116), and the VicarVision database for face recognition (45) and 200 negative from the Corel Gallery. This data set was randomly divided into a 50% train and test set. Faces were chosen because it was in line with our application goal. Moreover, it is a domain that lends itself to qualitative interpretation of evolving pattern prototypes.

The positive data set is used to discover the micro, meso, and macro patterns. Next, the positive and negative sets are used to evaluate the discovered macro patterns through diverse density and to determine the threshold value for a macro pattern. The resulting macro patterns are used to classify images from the test set. We will use accuracy, confusion matrices, and precision-recall curves to evaluate the quality of the predictions.

For the representation of micro patterns, we used 16 buckets for H, 6 buckets for I, and the 3 features for texture; see also Section II.A. We used a structured approach to determine the optimal values for sample frequencies, classification algorithms, threshold values, cluster precision, and the simulated

Table 1. Impact of pattern size on performance

Level	Size	Performance	p
Micro	8×8	76% ± 3.5	> 0.1
	12×12	71% ± 5.0	
Meso	1×1	62% ± 2.0	< 0.05
	2×2	75% ± 3.5	
	3×3	74% ± 2.6	
	4×4	79% ± 3.5	
Macro	2	74% ± 2.9	> 0.1
	4	77% ± 3.2	
	6	73% ± 4.7	
	8	78% ± 2.0	
	10	76% ± 4.0	

annealing parameters. The simulated annealing parameters were selected such that they were guaranteed to give the same results over 10 runs. The following values were determined for the three parameters: start temperature: 10, end temperature: 0.01, and cooling factor: 0.995. The sampling frequency and threshold value were increased with steps of respectively 100 and 0.1, until performance was optimal. Because sampling is non deterministic, we ran all experiments three times and used Repeated Measures ANOVA to test for significant differences in performance.

B. Impact of Pattern Size

Table 1 provides an overview of the results for the impact of pattern size at the various levels. For micro patterns, only at the larger pattern sizes differences in texture between the micro pattern prototypes evolve. Figure 2 gives an example of a face image reconstructed with face micro patterns of different size, and the reconstruction of a non face image using face micro patterns. Image areas that exceed the distance threshold for the closest pattern prototype are not allocated to a cluster - the transparent areas.

However, there is a tradeoff as smaller pattern sizes lead to better performance, though not significant. This is in contrast with the original Visual Alphabet setting classifier results: 16×16 patterns outperform 8×8 patterns [9]. Such a result, if significant, could be explained by the differences in outcome classes in terms of the concepts 'things' and 'stuff', as mentioned in the introduction. Things (e.g., objects, people, faces) consist of stuff; however, settings (e.g., a beach scene, countryside) typically consist of a small number of relatively large areas of stuff, whereas things consist of a large number of small areas of stuff [1], [6], [12], [15], [20].

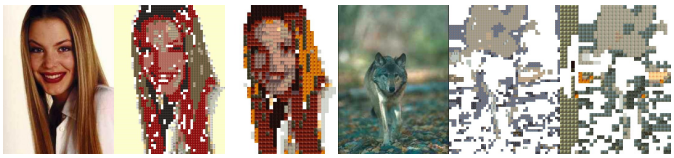
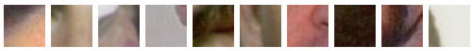
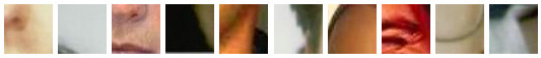


Fig. 2. Impact of pattern size: examples of a positive image (left) and a negative image (right) represented in terms of the micro pattern prototypes (cluster centers) with increasing pattern size.

Table 2. Impact of pattern size: meso pattern prototypes

Size	Meso Pattern Prototypes
1×1	
2×2	
3×3	
4×4	

We experimented with the size of the meso patterns. Please recall that meso patterns are patterns of micro level prototypes. A visual overview of varying pattern sizes is provided in Table 2. For larger patterns, we can clearly see meso pattern prototypes evolve, which are specific for faces; i.e., nose, mouth, neck, eye, chin, and borders between head and background. The 3×3 and 4×4 pattern results are significantly better than 1×1 pattern results (Table 1). This also demonstrates that classification on the basis of spatial configurations of micro patterns (size: $> 1 \times 1$) performs better than on micro patterns alone. This is a key result as it justifies the use of the intermediate meso level. Moreover, support for such an intermediate level can be found in literature [2], [21].

At the macro level, pattern size can be interpreted as the number of meso pattern prototypes to be used in a macro pattern to optimize classification performance. We experimented with $2, 4, 6, 8,$ and 10 meso pattern prototypes. From visual inspection, we noted that with more than 4 prototypes, features started to appear that were not unique to faces. However, note that for classification we only look at the distance to the best matching meso pattern prototype in the macro pattern.

As is shown in Table 1, no significant differences in performance for the number of meso pattern prototypes in a macro pattern were found. However, there are large differences in computational cost.

C. Impact of the Distance Measure

The distance measure of choice is another key parameter, with which was experimented. Euclidean, normalized correlation, and Mahalanobis distance were applied at the various levels; see also Equations 2 and 3 in Section II.A. The results of each of the distance measures on all levels can be found in Table 3.

For micro patterns, the difference between Euclidean and normalized correlation is not significant ($p > 0.1$); however, the difference between Mahalanobis and the other measures is ($p < 0.5$). Also for meso patterns, using Mahalanobis provided the best results; however, this is only significant for the

Table 3. Impact of distance measure on performance

Level	Distance	Performance	p
Micro	Euclidean	74% ± 2.1	< 0.05
	norm. correlation	75% ± 3.2	
	Mahalanobis	86% ± 2.1	
Meso	Euclidean	71% ± 5.3	< 0.05
	norm. correlation	77 ± 4.4	
	Mahalanobis	80% ± 1.7	
Macro	Euclidean	75% ± 6.0	< 0.05
	norm. correlation	62% ± 9.5	
	Mahalanobis	-	

comparison between Mahalonobis and Euclidean ($p < 0.5$). Mahalanobis distance essentially weights the features so the results indicate this is important for this class of images, and more important for micro than for meso level. A reason could be that micro patterns' H, I, and texture features correlation among each other, in contrast to the meso pattern, which is a pattern of micro pattern clusters.

For macro patterns, the results are different. The Mahalanobis distance could not be calculated because the feature covariance matrix was singular. Euclidean distance outperforms normalized correlation in this case ($p < 0.05$). A possible explanation for this is that macro patterns are essentially binary, whereas meso and micro patterns contain histogram and frequency information respectively. So, correlations are less likely to occur and the value add of normalization is smaller for macro patterns.

D. Impact of the Number of Pattern Prototypes

The final variable under investigation was the number of pattern prototypes; i.e., clusters. Table 4 provides the results. See Figure 3 for examples of reconstructed images with increasing number of clusters; image representation seems to improve with larger numbers of clusters. However, the experiments actually show that using 5 clusters gives optimal results (significant over 10 and 20 clusters at the 0.05 level). So, the downside of an increase in ambiguity outweighs the benefits of a closer image representation. For meso patterns, a similar tradeoff is shown. The results for $5, 10$, and 15 clusters are significantly better than the results for 20 clusters ($p < 0.5$). Similar as in the pattern sizes experiments, patterns evolve that correspond to useful salient features for recognizing faces; e.g., pieces of noses, hair, and mouth. At the macro level, all experiments result in a single pattern.

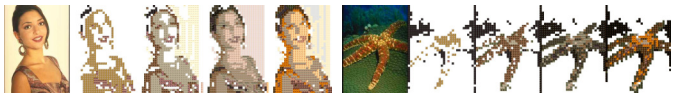


Fig. 3. Impact of number of prototypes: examples of a positive image (left) and a negative image (right) represented in terms of the micro pattern prototypes (cluster centers) with increasing number of clusters.

Table 4. Impact of the number of pattern prototypes on performance

Level	Clusters	Performance	p
Micro	5	83% ± 2.9	< 0.05
	10	73% ± 2.1	
	15	80% ± 1.7	
	20	75% ± 4.2	
Meso	5	71% ± 5.5	< 0.05
	10	77% ± 1.2	
	15	75% ± 3.8	
	20	64% ± 3.6	

E. Experimental Comparison

Finally, we ran an experiment with the optimal settings as discussed in the previous sections. Some exceptions were made: for 8×8 pixel size prototypes at the micro level the covariance matrix was singular; so, we used Euclidean rather than Mahalanobis distance at the meso level. See Table 5 for a confusion matrix, presenting the average results over all runs. The overall accuracy was 86%. The optimal threshold on the score was determined in training data only.

Our method is in essence an unsupervised, one class learning technique, only in the final macro stage, information from negative images is used to create a classifier. Therefore, we compare it to other similar methods for detecting image content in images that also exploit spatial information:[2], [12], [21]. We will summarize their approach and provide benchmark results. Note that this comparison is for illustration purposes only, given the differences in data sets across all methods.

In [21], areas are localized with a lot of structure and representations of these areas are clustered with a k-means algorithm. A greedy search algorithm is used to further extract the best set of prototypes. This results in 90% accuracy for frontal faces and 87% for car images (taken from behind). In [2], an approach similar to Weber et al is followed. However, they also took the geometric relations between patterns into account. They report 85% accuracy for car images (taken from the side). In [12], images are classified into a fixed number of categories and represented image content by average RGB values. This approach was tested on natural settings such as mountains, fields, and waterfalls and evaluated using precision-recall curves. See Figure 4 for a comparison. To ease the latter comparison, we have downscaled this study's results to the precision of [12], at full recall.

IV. Conclusion

In this chapter, a hierarchical multi-level Visual Alphabet approach for image representation and classification has been introduced. This method is inspired by visual perception theory

Table 5. Confusion matrix

		Classification	
		+	-
Image	+	81%	19%
	-	10%	90%

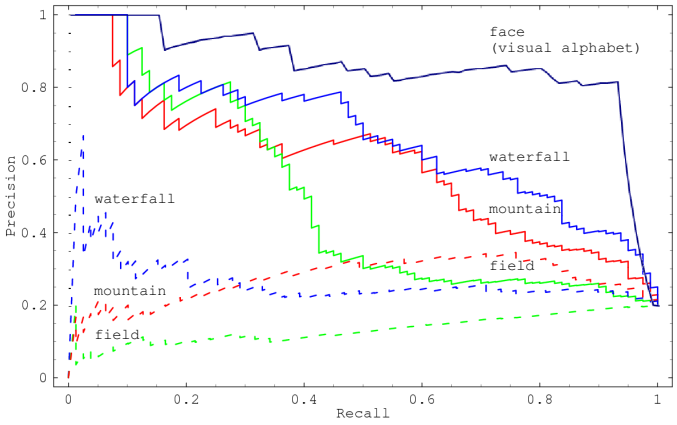


Fig. 4. Precision-recall curve (categories: waterfall, mountains, fields), adapted from [12] and the Visual Alphabet method (category: faces). Dashed lines are precision and recall curves on global histogram only for image classes from [12]. Visual Alphabets results scaled down to a 0.2 precision at 100% recall, to simplify a fair comparison.

and aims to integrate key concepts from opposing traditions. The representation is hierarchical; although, abstract, generic, and computationally complex features are avoided. The micro and meso level features are derived directly from the positive examples only. The multi-level Visual Alphabet method is a well suited, generic, and coarse method for the representation of image content. It covers most of the continuum between stuff and things. The aim of the additional meso level is to primarily capture what micro pattern prototypes co-occur spatially.

The initial results are encouraging. A high level comparison is provided with the classification results from [12], among others. Moreover, the results confirmed that useful, salient features are being evolved at the intermediate meso level. From a qualitative point of view, the meso patterns seem to match with patterns to be expected from a classifier, whose task it is to recognize faces. From a quantitative point of view, the pattern size experiments at the meso level have confirmed that representations that go beyond simple distributions of micro patterns (i.e., meso patterns larger than a single micro pattern prototype) provide significantly better results. In other words, adding the meso level actually adds value over our prior, two level approach, towards visual alphabets.

Image representation and classification is again successfully achieved using Visual Alphabets. This article presented a significant extension to the original Visual Alphabet method [8], [9]. It is unique in that it merged two opposing paradigms in human visual perception. The experiments prove both quantitatively and qualitatively that the additional middle layer adds value, and provide guidance for understanding and setting the main parameters. Given its theoretical framework and these initial results, this multi-level approach to Visual Alphabets may be a promising method to narrow the semantic gap [4], and applicable to a wider set of problems than the original approach.

REFERENCES

[1] A. Agarwal and B. Triggs. Multilevel image coding with hyperfeatures. *International Journal of Computer Vision*, 78(1):15–27, 2008.

[2] S. Agarwal and D. Roth. Learning a sparse representation for object detection. *Lecture Notes in Computer Science (Computer Vision)*, 2353:97–101, 2002.

[3] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In J.B. Bocca, M. Jarke, and C. Zaniolo, editors, *VLDB*, pages 487–499. Morgan Kaufmann, 1994.

[4] Y. Avrithis, N. E. O'Connor, S. Staab, and R. Troncy. Introduction to the special issue on "semantic multimedia". *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 6(2):137–138, 2008.

[5] A. Bosch, A. Zisserman, and X. Munoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008.

[6] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, [in press].

[7] J. J. Gibson. *The Ecological Approach to Visual Perception*. Boston, MA, USA: Houghton Mifflin, 1979.

[8] M. Israël, E. L. van den Broek, P. van der Putten, and M. J. den Uyl. Automating the construction of scene classifiers for content-based video retrieval. In L. Khan and V. A. Petrushin, editors, *Proceedings of the Fifth ACM International Workshop on Multimedia Data Mining (MDM/KDD'04)*, pages 38–47, Seattle, WA, USA, 2004.

[9] M. Israël, E. L. van den Broek, P. van der Putten, and M. J. den Uyl. *Visual Alphabets: Video Classification by End Users*, chapter 10 (Part III: Multimedia Data Indexing and Retrieval), pages 185–206. Springer-Verlag: Berlin - Heidelberg, 2007.

[10] M. Lillholm and L. D. Griffin. Novel image feature alphabets for object recognition. In *9th International Conference on Pattern Recognition (ICPR 2008), December 8-11, 2008, Tampa, Florida, USA*, pages 1–4, 2008.

[11] O. Maron. *Learning from Ambiguity*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), 1998.

[12] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *Proceedings Fifteenth International Conference on Machine Learning ICML-98*, pages 341–349. Morgan Kaufmann, 1998.

[13] D. Marr. *Vision*. New York, NY, USA: W. H. Freeman and Co., 1982.

[14] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[15] R. W. Picard and T. P. Minka. Vision texture for annotation. *Multimedia Systems*, 3(1):3–14, 1995.

[16] M. Sonka, V. Hlavac, and R. Boyle. *Image processing, analysis and machine vision*. San Francisco, CA, USA: PWS publishing, 1999.

[17] M. To, P. G. Lovell, T. Troscianko, and D. J. Tolhurst. Summation of perceptual cues in natural visual scenes. *Proceedings of the Royal Society B*, 275(1649):2299–2308, 2008.

[18] E. L. van den Broek, P. M. F. Kisters, and L. G. Vuurpijl. Content-based image retrieval benchmarking: Utilizing color categories and color distributions. *Journal of Imaging Science and Technology*, 49(3):293–301, 2005.

[19] E. L. van den Broek, Th. E. Schouten, and P. M. F. Kisters. Modeling human color categorization. *Pattern Recognition Letters*, 29(8):1136–1144, 2008.

[20] J. C. van Gemert, C. G. M. Snoek, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Comparing compact codebooks for visual categorization. *Computer Vision and Image Understanding*, 114(4):450–462, 2010.

[21] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. *Lecture Notes in Computer Science (Computer Vision)*, 1842:18–32, 2000.

[22] S. Zhu, C. Guo, Y. Wang, and Z. Xu. What are textons? *Int. J. Comput. Vision*, 62(1-2):121–143, 2005.