A Comparison of Color Features for Visual Concept Classification

Koen E.A. van de Sande
ISLA, Informatics Institute
University of Amsterdam
Kruislaan 403, 1098SJ
Amsterdam, The Netherlands
ksande@science.uva.nl

Theo Gevers
ISLA, Informatics Institute
University of Amsterdam
Kruislaan 403, 1098SJ
Amsterdam, The Netherlands
gevers@science.uva.nl

Cees G.M. Snoek
ISLA, Informatics Institute
University of Amsterdam
Kruislaan 403, 1098SJ
Amsterdam, The Netherlands
cgmsnoek@science.uva.nl

ABSTRACT

Concept classification is important to access visual information on the level of objects and scene types. So far, intensity-based features have been widely used. To increase discriminative power, color features have been proposed only recently. As many features exist, a structured overview is required of color features in the context of concept classification.

Therefore, this paper studies 1. the invariance properties and 2. the distinctiveness of color features in a structured way. The invariance properties of color features with respect to photometric changes are summarized. The distinctiveness of color features is assessed experimentally using an image and a video benchmark: the PASCAL VOC Challenge 2007 and the Mediamill Challenge.

Because color features cannot be studied independently from the points at which they are extracted, different point sampling strategies based on Harris-Laplace salient points, dense sampling and the spatial pyramid are also studied.

From the experimental results, it can be derived that invariance to light intensity changes and light color changes affects concept classification. The results reveal further that the usefulness of invariance is concept-specific.

Categories and Subject Descriptors

I.4.7 [**Image Processing and Computer Vision**]: Feature Measurement; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Performance, Measurement

Keywords

Color, Invariance, Concept Classification, Object and Video Retrieval, Bag-of-Features, Spatial Pyramid

CIVR'08, July 7–9, 2008, Niagara Falls, Ontario, Canada.
Copyright 2008 ACM 978-1-60558-070-8/08/07 ...$5.00.

1. INTRODUCTION

Image concept classification is important to access visual information on the level of objects (buildings, cars, *etc.*) and scene types (outdoor, vegetation, *etc.*). In general, systems in both image retrieval [12, 17, 30] and video retrieval [3, 9, 24, 28] use machine learning based on image descriptions to distinguish object and scene concepts. However, there can be large variations in lighting and viewing conditions for real-world scenes, complicating the description of images. A change in viewpoint will yield shape variations such as the orientation and scale of the object in the image plane. Therefore, effective visual classification methods should be invariant to accidental recording circumstances.

Among all invariant feature extraction methods on offer, the one based on salient point detection has gained widespread acceptance in both the computer vision and video retrieval community [9, 28, 30]. Salient point detection methods and corresponding region descriptors can robustly detect regions which are translation-, rotation- and scale-invariant. Hereby addressing the problem of viewpoint changes [15, 19]. However, changes in the illumination of a scene can greatly affect the performance of object recognition if the descriptors used are not robust to these changes. To increase illumination invariance and discriminative power, color features have been proposed [1, 7, 20, 27]. However, as there are many different color models, a comparison is required based on their illumination invariance properties and their distinctiveness in the context of concept classification.

Because color features are often computed around specific salient points, they cannot be evaluated independently of the point sampling strategy used. The sampling strategy can have a profound effect on 1. the discriminative power and 2. the computational efficiency of color features. A salient point detector is more efficient than a set of points densely sampled over the image grid, but has less discriminative power. Similarly, the extension of point sampling to multiple image areas, the spatial pyramid [12], adds discriminative power, at the expense of additional computational effort.

This paper compares 1. the invariance properties and 2. the distinctiveness of color features in a structured way. The *invariance properties* of color features with respect to photometric changes are summarized. The *distinctiveness* of color features is analyzed experimentally using two representative and widely-adopted benchmarks from the image domain and the video domain. The benchmarks are very different in nature: the image benchmark PASCAL VOC Challenge 2007 [4] consists of photographs and the Mediamill Challenge [25] consists of news broadcast videos.

This paper is organized as follows. In section 2, the relation of this paper with other work is discussed. In section 3, an overview is

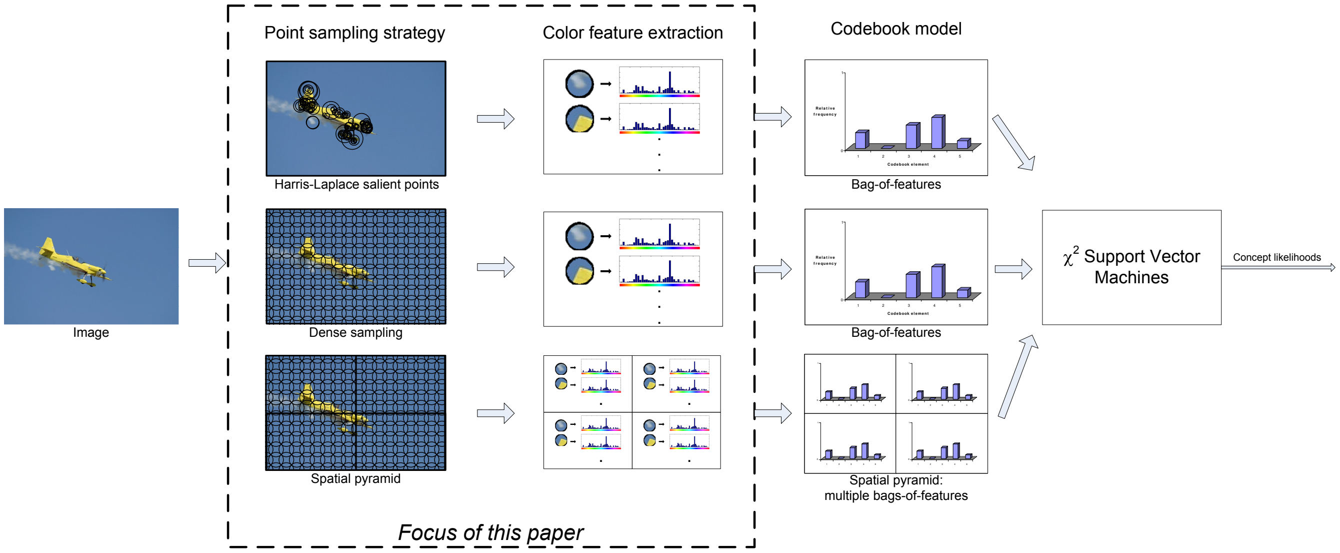

Figure 1: Overview of concept classification using the codebook model. In the first stage, points are sampled in the image, using either Harris-Laplace or dense sampling. In the color feature extraction stage, color features are extracted around every sampled point. Next, the color features of an image, the 'bag-of-features', are vector-quantized using a codebook. This forms the input to the SVM classifier, which outputs a concept likelihood score for the image. The spatial pyramid divides the image into 1x1, 2x2, 4x4, *etc*. regions. For every region, the color features extracted from that region are vector-quantized. In effect, every image region has its own bag-of-features. These are then combined at the learning stage. The focus of this paper lies on the point sampling strategy and the color features.

given of color features and their invariance properties. The experimental setup is presented in section 4. In section 5, a discussion of the results is given. Finally, in section 6, conclusions are drawn.

2. RELATED WORK

Many current systems for concept classification use the bag-of-features model as a basic building block. This model vector-quantizes local features. It is also referred to as 'textons' [14], 'object parts' [6], 'visual words' [23] and 'codebooks' [10, 13]. Figure 1 gives an overview of the components of concept classification based on the codebook model. The first component is the strategy to sample points for the local features. Other important components are the color features which describe the point, the choice of visual codebook and the machine learning algorithm used.

2.1 Point sampling strategy

Local features are often extracted at either salient points [6, 13, 30] or densely sampled over the image grid [5, 16]. For salient point extraction, Zhang [30] observes that the Harris-Laplace and Laplacian detectors are the preferable choice in terms of classification accuracy. This detector uses the Harris corner detector to find potential scale-invariant points. It then selects a subset of these points for which the Laplacian-of-Gaussians reaches a maximum over scale. Dense sampling [10] is a uniform sampling over the image grid with a fixed pixel interval between the points. In the context of concept classification, a distinction is made between two classes of concepts: object classification and scene type classification. Dense sampling has been shown to be advantageous for scene type classification, since salient points do not capture the entire appearance of an image. For object classification, salient points can be advantageous because they ignore homogenous areas in the image. If the object background is not highly textured, then most salient points will be located on the object or the object boundary.

In conclusion, to prevent a bias towards concept classes, *i.e*. objects or scene type, both salient point sampling and dense sampling are evaluated to assess concept classification accuracy.

2.2 Color features

For point description, the SIFT feature by Lowe [15] is generally used because of its good classification accuracy [18, 30]. The SIFT feature captures the local shape around a point using edge orientation histograms. However, SIFT operates on intensity information only, ignoring the color information available. Van de Weijer [27] and Bosch [1] have recognized this weakness and have proposed HueSIFT and HSV-SIFT, respectively. Also, there are color histograms in many color models, all with different levels of invariance. In [26], we performed an analysis of the invariance properties for these and other color features, such as color histograms, color moment invariants [20] and color extensions of SIFT.

More attention is needed to discuss color features and their invariance properties, therefore these will be further covered in section 3.

2.3 Codebook-based classification

Codebooks for bag-of-features methods are usually constructed by clustering the features of points from a set of training images using a methods such as k-means. Other clustering methods, for example radius-based clustering [10], are known to improve performance. A more significant aspect of bag-of-features methods, however, is the codebook size [22]. For example, this was observed by Jiang [9] in the context of video concept classification.

To improve performance further, Lazebnik [12] proposed the spatial pyramid, which includes spatial information into the bag-of-features model. The spatial pyramid divides the image into 1x1, 2x2, 4x4, *etc*. regions. For every region, the features extracted from that region are vector-quantized. In effect, every image region is an image in itself. These are then combined using a weighting

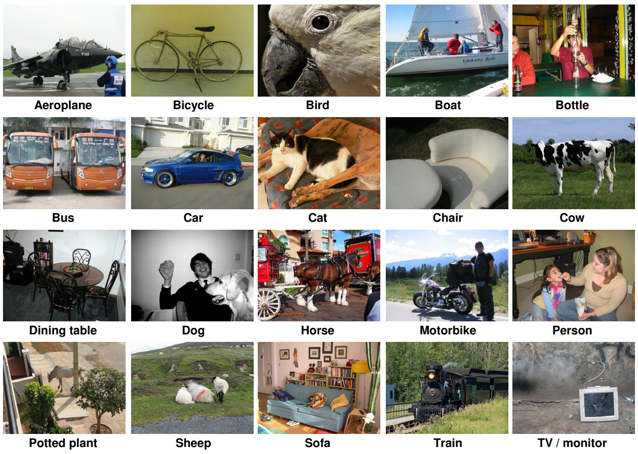

Figure 2: Concepts of the PASCAL Visual Object Challenge 2007, used in the image benchmark of experiments 1 and 2.

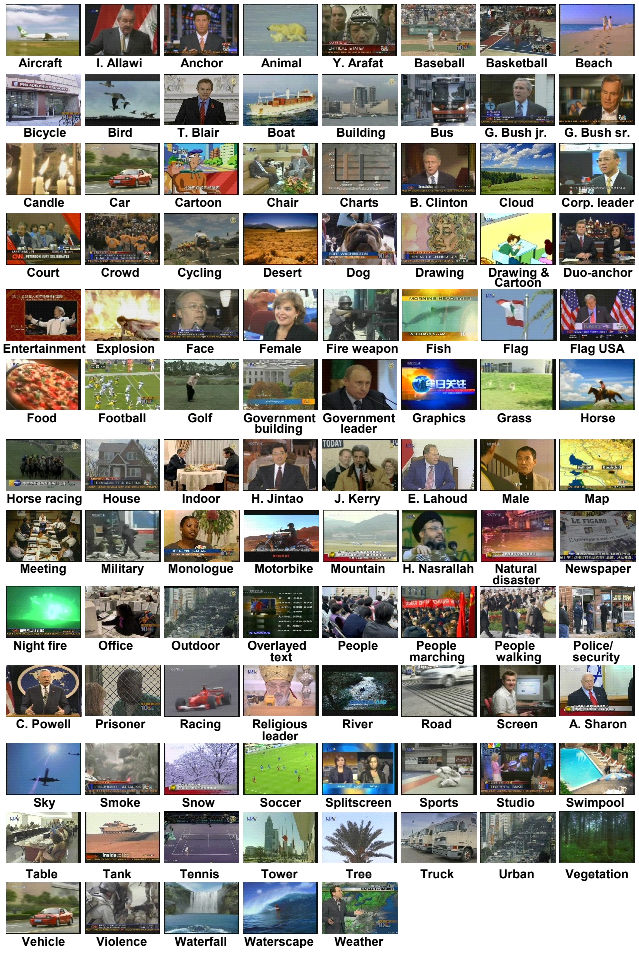

Figure 3: Concepts of the Mediamill Challenge, used in the video benchmark of experiment 3. Image based on [25].

scheme which depends on the level in the spatial pyramid. Results from [12] show that the first division into four image quarters is most significant w.r.t. performance.

For learning concept classifiers, the Support Vector Machines (SVM) algorithm is used by all state-of-the-art systems. Variations in classification accuracy are possible due to the choice of SVM kernel function. Zhang [30], Wang [28] and Jiang [9] observe that the χ^2 SVM kernel is one of the best kernel functions for concept classification. Zhang additionally observes that the Earth-Movers Distance kernel provides similar performance to the χ^2 kernel, but is much more expensive to compute. The χ^2 SVM kernel is based on the χ^2 distance between two feature vectors $\vec{F}$ and $\vec{F'}$:

$$d(\vec{F}, \vec{F'}) = \frac{1}{2} \sum_{i=1}^{n} \frac{(\vec{F_i} - \vec{F'_i})^2}{\vec{F_i} + \vec{F'_i}}$$

For notational convenience, $\frac{0}{0}$ is assumed to be equal to 0 iff $\vec{F_i} = \vec{F'_i} = 0$.

In conclusion, in this paper we take the most promising components of concept classification based on the codebook model as starting point: a large codebook in combination with the efficient χ^2 SVM kernel. The effect of extending the bag-of-features model with a spatial pyramid will be studied in the experiments.

2.4 Evaluation

To evaluate concept classification, there are multiple benchmarks to chose from. For example, Caltech-256 [8] and the PASCAL VOC Challenge [4] provide a large set of images with annotated concepts. The Caltech-256 dataset consists of 30,607 images from 256 different concepts. The PASCAL VOC 2007 dataset consists of 9,963 images from 20 different concepts. However, the advantage of the PASCAL VOC 2007 dataset is that multiple concepts have been annotated per image, if they are present. For the Caltech-256 dataset, this is not the case. This paper uses the PASCAL VOC 2007 dataset as an image benchmark. Its concepts are illustrated in figure 2.

For video, most benchmarks available are based on TRECVID [24] video data. The Mediamill Challenge [25] provides a baseline and annotations for 101 concepts. The Columbia374 [29] provides a baseline for 374 concepts based on LSCOM concept annotations [21]. Both are based on the same TRECVID data of news broadcasts from English, Arabic and Chinese TV channels. However, the Mediamill Challenge defines a repeatable experiment to evaluate the performance of visual features only. Therefore, this paper uses the Mediamill Challenge as a video benchmark. Its concepts are illustrated in figure 3.

Systems with the best performance in image retrieval [17] and video retrieval [28] use combinations of multiple features for concept classification. The basis for these combinations is formed by good individual features and multiple point sampling strategies. Therefore, for a state-of-the-art comparison, color features should be studied in combination with a good point sampling strategy.

In conclusion, the point sampling strategy and color features will be evaluated on real-world image and video benchmarks in a state-of-the-art environment, as shown in figure 1.

3. COLOR FEATURES

In this section, color features are presented and their invariance properties are summarized. First, color features based on histograms are discussed. Then, color moments and color moment invariants are presented. Finally, color features based on SIFT are detailed. See table 1 for an overview of features and their invariance properties.

	Light intensity change	**Light intensity shift**	**Light intensity change and shift**	**Light color change**	**Light color change and shift**
RGB Histogram	-	-	-	-	-
O_1, O_2	-	+	-	-	-
O_3, Intensity	-	-	-	-	-
Hue	+	+	+	-	-
Saturation	+	+	+	-	-
r, g	+	-	-	-	-
Transformed color	+	+	+	+	+
Color moments	-	+	-	-	-
Moment invariants	+	+	+	+	+
SIFT (∇I)	+	+	+	+	+
HSV-SIFT	+	+	+	+/-	+/-
HueSIFT	+	+	+	+/-	+/-
OpponentSIFT	+/-	+	+/-	+/-	+/-
W-SIFT	+	+	+	+/-	+/-
rgSIFT	+	+	+	+/-	+/-
Transf. color SIFT	+	+	+	+	+

Table 1: Invariance of features (section 3) against types of lighting changes. Invariance is indicated with '+', lack of invariance is indicated with '-'. A '+/-' indicates that the intensity SIFT part of the feature is invariant, but the color part is not. All color features can be selected in the color feature extraction stage from figure 1.

In short, the photometric changes and their corresponding invariance properties are:

- Light intensity changes include shadows and lighting geometry changes such as shading. When a feature is invariant to light intensity changes, it is *scale-invariant* with respect to (light) intensity.
- Light intensity shifts correspond to object highlights under a white light source and scattering of a white source. When a feature is invariant to a light intensity shift, it is *shift-invariant* with respect to light intensity.
- Light intensity change and shift allows combinations of the above two conditions. When a feature is robust to these changes is scale-invariant and shift-invariant with respect to light intensity.
- Light color change corresponds to a change in the illumination color and light scattering, amongst others.
- Light color change and shift corresponds to changes in the illumination, as above, and to object highlights under an arbitrary light source.

For additional details and derivations, we refer to [26].

3.1 Histograms

RGB histogram The RGB histogram is a combination of three 1-D histograms based on the R, G and B channels of the RGB color space. This histogram possesses no invariance properties, see table 1.

Opponent histogram The opponent histogram is a combination of three 1-D histograms based on the channels of the opponent color space:

$$\begin{pmatrix} O1 \\ O2 \\ O3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix}. \quad (1)$$

The intensity is represented in channel $O3$ and the color information is in channels $O1$ and $O2$. Due to the subtraction in $O1$ and $O2$, the offsets will cancel out if they are equal for all channels (e.g. a white light source). Therefore, these color models are shift-invariant with respect to light intensity. The intensity channel $O3$ has no invariance properties. The histogram intervals for the opponent color space have ranges different from the RGB model.

Hue histogram In the HSV color space, it is known that the hue becomes unstable around the grey axis. To this end, Van de Weijer *et al.* [27] apply an error analysis to the hue. The analysis shows that the certainty of the hue is inversely proportional to the saturation. Therefore, the hue histogram is made more robust by weighing each sample of the hue by its saturation. The H and the S color models are scale-invariant and shift-invariant with respect to light intensity.

rghistogram In the normalized RGB color model, the chromaticity components r and g describe the color information in the image (b is redundant as $r + g + b = 1$):

$$\begin{pmatrix} r \\ g \\ b \end{pmatrix} = \begin{pmatrix} \frac{R}{R+G+B} \\ \frac{G}{R+G+B} \\ \frac{B}{R+G+B} \end{pmatrix}. \quad (2)$$

Because of the normalization, r and g are scale-invariant and thereby invariant to light intensity changes, shadows and shading [26].

Transformed color distribution A RGB histogram is not invariant to changes in lighting conditions. However, by normalizing the pixel value distributions, scale-invariance and shift-invariance is achieved with respect to light intensity. Because each channel is normalized independently, the feature is also normalized against changes in light color and arbitrary offsets:

$$\begin{pmatrix} R' \\ G' \\ B' \end{pmatrix} = \begin{pmatrix} \frac{R-\mu_R}{\sigma_R} \\ \frac{G-\mu_G}{\sigma_G} \\ \frac{B-\mu_B}{\sigma_B} \end{pmatrix}, \quad (3)$$

with μ_C the mean and σ_C the standard deviation of the distribution in channel C. This yields for every channel a distribution with $\mu = 0$ and $\sigma = 1$.

3.2 Color moments and moment invariants

A color image corresponds to a function I defining RGB triplets for image positions (x, y): $I : (x, y) \mapsto (R(x, y), G(x, y), B(x, y))$. By regarding RGB triplets as data points coming from a distribution, it is possible to define moments. Mindru *et al.* [20] have defined *generalized color moments* M_{pq}^{abc}:

$$M_{pq}^{abc} = \int\int x^p y^q [I_R(x,y)]^a [I_G(x,y)]^b [I_B(x,y)]^c dxdy.$$

M_{pq}^{abc} is referred to as a generalized color moment of *order* $p+q$ and *degree* $a+b+c$. Note that moments of order 0 do not contain any spatial information, while moments of degree 0 do not contain any photometric information. Thus, moment descriptions of order 0 are rotationally invariant, while higher orders are not. A large number of moments can be created with small values for the order and degree. However, for larger values the moments are less stable. Typically generalized color moments up to the first order and the second degree are used.

By using the proper combination of moments, it is possible to normalize against photometric changes. These combinations are called *color moment invariants*. Invariants involving only a single color channel (*e.g.* out of a, b and c two are 0) are called 1-band invariants. Similarly there are 2-band invariants involving only two out of three color bands. 3-band invariants involve all color channels, but these can always be created by using 2-band invariants for different combinations of channels.

Color moments The color moment feature uses all generalized color moments up to degree 2 and order 1. This lead to nine possible combinations for the degree: M_{pq}^{000}, M_{pq}^{100}, M_{pq}^{010}, M_{pq}^{001}, M_{pq}^{200}, M_{pq}^{110}, M_{pq}^{020}, M_{pq}^{011}, M_{pq}^{002} and M_{pq}^{101}[†]. Combined with three possible combinations for the order: M_{00}^{abc}, M_{10}^{abc} and M_{01}^{abc}, the color moment feature has 27 dimensions. These color moments only have shift-invariance. This is achieved by Mindru by subtracting the average in all input channels before computing the moments.

Color moment invariants Color moment invariants can be constructed from generalized color moments. All 3-band invariants are computed from Mindru *et al.* [20]. To be comparable, the $\tilde{C}_{02}$ invariants are considered. This gives a total of 24 color moment invariants, which are invariant to all the properties listed in table 1.

3.3 Color SIFT features

SIFT The SIFT feature proposed by Lowe [15] describes the local shape of a region using edge orientation histograms. The gradient of an image is shift-invariant. Under light intensity changes, *i.e.* a scaling of the intensity channel, the gradient direction and the relative gradient magnitude remain the same. Because the SIFT feature is normalized, the gradient magnitude changes have no effect on the final feature. Light color changes have no effect on the feature because the input image is converted to gray-scale, after which the intensity scale-invariance argument applies. To compute SIFT features, the version described by Lowe [15] is used.

HSV-SIFT Bosch [1] computes SIFT features over all three channels of the HSV color model, instead of over the intensity channel only. This gives 3x128 dimensions per feature, 128 per channel. Drawback of this approach is that the periodicity in the hue channel is not addressed. Moreover, the instability of the hue for low saturation is ignored.

The properties of the H and the S channels also apply to this feature: it is scale-invariant and shift-invariant. However, the H and the S SIFT features are not invariant to light color changes; only the intensity SIFT feature (V channel) is invariant to this. Therefore, the feature is only partially invariant to light color changes.

HueSIFT Van de Weijer [27] introduces a concatenation of the hue histogram (see section 3.1) with the SIFT feature. When compared to HSV-SIFT, the usage of the weighed hue histogram addresses the instability of the hue around the grey axis. Because the bins of the hue histogram are independent, there are no problems with the periodicity of the hue channel for HueSIFT. Similar to the hue histogram, the HueSIFT feature is scale-invariant and shift-invariant. However, only the SIFT feature is component of this feature is invariant to illumination color changes or shifts; the hue histogram is not.

OpponentSIFT OpponentSIFT describes all the channels in the opponent color space (eq. (1)) using SIFT features. The information in the O_3 channel is equal to the intensity information, while the other channels describe the color information in the image. However, these other channels do contain some intensity information: hence they are not invariant to changes in light intensity.

W-SIFT In the opponent color space (eq. (1)), the O_1 and O_2 channels still contain some intensity information. To add invariance to intensity changes, [7] proposes the W invariant which eliminates the intensity information from these channels. The W-SIFT feature uses the W invariant, which can be defined for the opponent color space as $\frac{O_1}{O_3}$ and $\frac{O_2}{O_3}$. Because of the division by intensity, the scaling in the diagonal model will cancel out, making W-SIFT scale-invariant with respect to light intensity. As for the other colorSIFT features, the color component of the feature is not invariant to light color changes.

***rg*SIFT** For the rgSIFT feature, descriptions are added for the r and g chromaticity components of the normalized RGB color model from eq. (2), which is already scale-invariant. Because the SIFT feature uses derivatives of the input channels, the rgSIFT feature becomes shift-invariant as well. However, the color part of the feature is not invariant to changes in illumination color.

Transformed color SIFT For the transformed color SIFT, the same normalization is applied to the RGB channels as for the transformed color histogram (eq. (3)). For every normalized channel, the SIFT feature is computed. The feature is scale-invariant, shift-invariant and invariant to light color changes and shift.

4. EXPERIMENTAL SETUP

Our implementation follows the general scheme for concept classification based on the codebook model, as detailed in section 2 and summarized in figure 1. In this section, we outline further details of the experimental setup used to evaluate the different color features. Then, the experiments and the two benchmarks used for evaluation are described: the PASCAL VOC Challenge 2007 and the Mediamill Challenge. After discussing these benchmarks and their datasets, evaluation criteria are given.

4.1 Implementation

To emperically test the color features, they are used inside local features based on either salient points [15, 30] or dense sampling. For salient point extraction, we choose the Harris-Laplace point detector [19]. For dense sampling, the sample distance used is 6 pixels. The color features from section 3 are computed over the area around the points. To achieve comparable features for different scales, all regions are proportionally resampled to a uniform patch size of 60 by 60 pixels.

To construct a fixed-length feature vector, the bag-of-features model with a visual codebook is used. The visual codebook is constructed using the k-means algorithm. The k-means algorithm is repeated 20 times on 125,000 features randomly drawn from the training set to obtain 4,000 clusters. Features from an image are assigned to the closest cluster. With $\vec{F}$ denoting the feature vector of length n, where n equals the codebook size:

$$\vec{F}_i = \frac{1}{m}\sum_{j=1}^{m}\psi(i,j)$$

where m is the number of features in the image and indicator function $\psi(i,j)$ equals 1 if the i^{th} cluster is closest to the j^{th} feature

[†] Because it is constant, the moment M_{pq}^{000} is excluded.

and 0 otherwise. Closeness is computed using the Euclidian distance. All elements of $\vec{F_i}$ are constrained to the range $[0, 1]$ by their definition.

The SVM algorithm is used to learn concept appearance models from feature vectors. Specifically, the LibSVM implementation [2] is used with a χ^2 kernel. For a spatial pyramid, there is a feature vector for every image section in the pyramid, *e.g.* the full image and the image quarters. The χ^2 distances for the different sections are normalized through division by the average distance. Then, the final distance is the weighted sum of the section distances. The weighting of the different image sections is performed as specified by Lazebnik [12]. For a spatial pyramid up to level 1, this is a weight of 1 for the full image and weights of $\frac{1}{4}$ for the image quarters.

4.2 Experiments

- *Experiment 1: Comparing point sampling strategies on image benchmark.*
  In this experiment, the performance of features using a standard bag-of-features codebook model is compared against the performance of features using a spatial pyramid up to level 1, *i.e.* the standard codebook plus the four image quarters. To rule out the effect of different sampling methods, the experiment is performed for both the Harris-Laplace detector and densely sampled points.

- *Experiment 2: Comparing color features on image benchmark.*
  In this experiment, the performance of color features is evaluated on an image benchmark. Based on the results of experiment 1, the spatial pyramid is used exclusively. Sampling methods used are the Harris-Laplace detector and dense sampling, or a combination of the two. The combination of sampling methods is constructed by concatenating the feature vectors. The color features used are listed in section 3 and table 1.

- *Experiment 3: Comparing color features on video benchmark.*
  In this experiment, the performance of color features is evaluated on a video benchmark. On the video benchmark, only the combination of the Harris-Laplace detector and dense sampling is used. The color features used are listed in section 3 and table 1.

4.3 Image benchmark

The PASCAL Visual Object Classes Challenge [4] provides a yearly benchmark for comparison of object classification systems. The PASCAL VOC Challenge 2007 dataset contains nearly 10,000 images of 20 different concepts (see figure 2), e.g. bird, bottle, car, dining table, motorbike and people. The dataset is divided into a predefined train set (5011 images) and test set (4952 images).

4.4 Video benchmark

The Mediamill Challenge by Snoek *et al* [25] provides an annotated video dataset, based on the training set of NIST TRECVID 2005 benchmark [24]. Over this dataset, repeatable experiments have been defined. The experiments decompose automatic category recognition into a number of components, for which they provide a standard implementation. This provides an environment to analyze which components affect the performance most. Since our features use visual information only, we focus on the visual experiment of the Challenge only. For this experiment, the Challenge provides a baseline performance.

Language	#	Source	Program	Length
Arabic	7	LBC	LBC Nahar	6h 46min
Arabic	5	LBC	LBC News (1pm)	2h 5min
Arabic	14	LBC	LBC News (8pm)	13h 34min
Chinese	13	CCTV4	Daily News	12h 19min
Chinese	11	CCTV4	News3	5h 5min
Chinese	10	NTDTV	NTD News (12pm)	4h 42min
Chinese	9	NTDTV	NTD News (7pm)	4h 15min
English	11	CNN	Aaron Brown	10h 42min
English	9	CNN	Live From	4h 11min
English	15	NBC	NBC Philadelphia	7h 5min
English	7	NBC	Nightly News	3h 18min
English	11	MSNBC	MSNBC News (11am)	5h 12min
English	15	MSNBC	MSNBC News (1pm)	7h 3min
Total	**137**			**86h 17min**

Table 2: Overview of the news broadcasts in the video benchmark.

The dataset of 86 hours is divided into a Challenge training set (70% of the data or 30993 shots) and a Challenge test set (30% of the data or 12914 shots). For every shot, the Challenge provides a single representative keyframe image. So, the complete dataset consists of 43907 images, one for every video shot. The dataset consists of television news from November 2004 broadcasted on six different TV channels in three different languages: English, Chinese and Arabic; see table 2 for a complete overview. On this dataset, the 101 concepts of the Mediamill Challenge are employed, listed in figure 3.

4.5 Evaluation criteria

The average precision is taken as the performance metric for determining the accuracy of ranked category recognition results, following the standard set in the PASCAL VOC Challenge 2007 and TRECVID. The average precision is a single-valued measure that is proportional to the area under a precision-recall curve. This value is the average of the precision over all shots judged relevant. Let $\rho^k = \{l_1, l_2, ..., l_k\}$ be the ranked list of items from test set A. At any given rank k, let $|R \cap \rho^k|$ be the number of relevant shots in the top k of ρ, where R is the set of relevant shots and $|X|$ is the size of set X. Average precision, AP, is then defined as:

$$AP(\rho) = \frac{1}{|R|} \sum_{k=1}^{|A|} \frac{|R \cap \rho^k|}{k} \psi(l_k) \qquad (4)$$

with indicator function $\psi(l_k) = 1$ if $l_k \in R$ and 0 otherwise. $|A|$ is the size of the answer set, *e.g.* the number of items present in the ranking.

When performing experiments over multiple object classes, the average precisions of the individual classes can be aggregated. This aggregation is called mean average precision (MAP). MAP is calculated by taking the mean of the average precisions. Note that MAP depends on the dataset used: scores of different datasets are not easily comparable.

5. RESULTS

5.1 Experiment 1: Comparing point sampling strategies on image benchmark

From the results shown in figure 4, it is observed that the spatial pyramid performs substantially better than the standard codebook model for all color features. This holds for both salient points

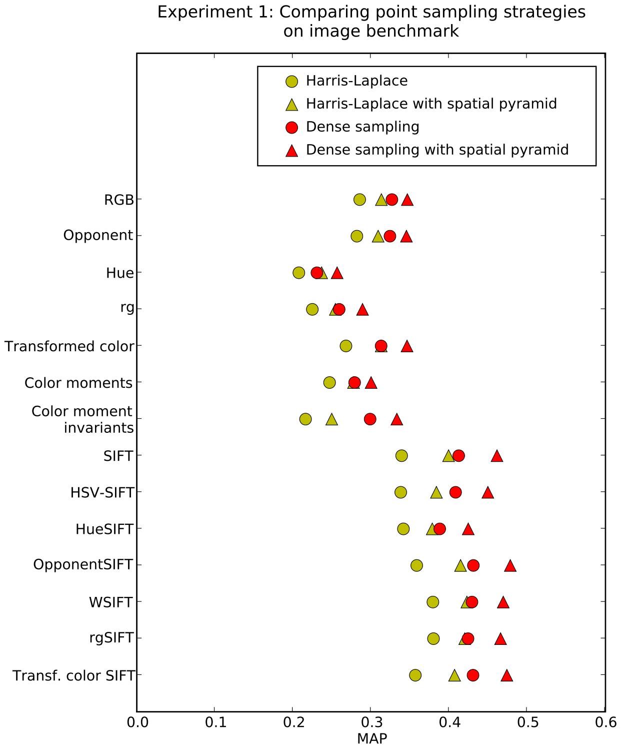

Figure 4: Evaluation of the standard codebook model and its extension, the spatial pyramid. Performance is measured on an image benchmark, the PASCAL VOC Challenge 2007, averaged over the 20 concepts from figure 2. Results are shown for both Harris-Laplace salient points and dense sampling.

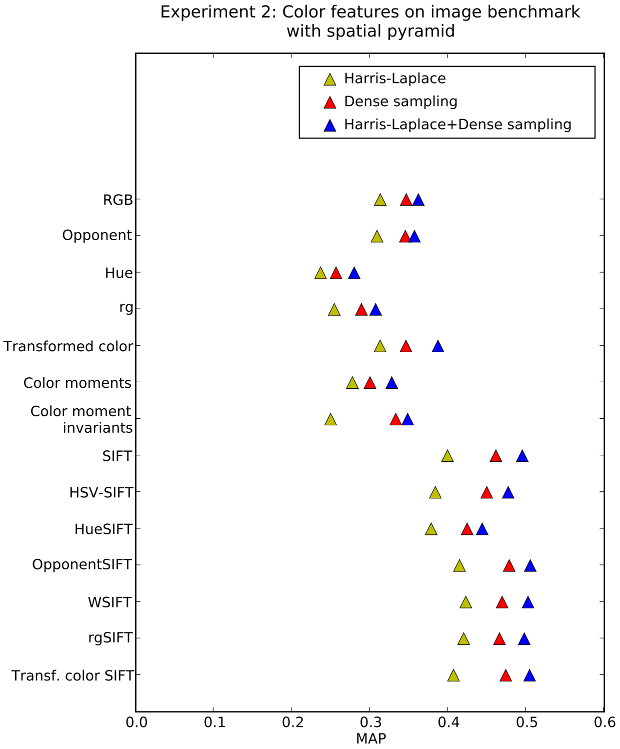

Figure 5: Evaluation of color features using either Harris-Laplace salient points, dense sampling, or both Harris-Laplace salient points and dense sampling. Performance is measured on an image benchmark, the PASCAL VOC Challenge 2007, averaged over the 20 concepts from figure 2.

from the Harris-Laplace detector and for densely sampled points. Adding spatial information to the codebook model provides important discriminative information for most concepts. However, for the concept aeroplane, detailed results (not shown) reveal there is no improvement or even a small degradation. This is explained by the lack of a fixed position for aeroplanes in photographs: they can occur in all image quarters. In that case, the spatial pyramid provides no benefits. Given that for all other concepts similar or better performance was obtained with the spatial pyramid, it is adopted for the remainder of the experiments.

Figure 4 also shows that, overall, dense sampling outperforms Harris-Laplace salient points. Even on a per-concept basis, Harris-Laplace is not convincingly better than dense sampling for specific concepts. This suggests that, if the computational resources are available, dense sampling should be the primary choice for point sampling.

5.2 Experiment 2: Comparing color features on image benchmark

The results from figure 5 show that using both the Harris-Laplace detector and densely sampled points clearly improves over the individual sampling strategies. On a per-concept basis, performance is either better than or has only minor differences from dense sampling. Therefore, to evaluate color features in a realistic setting, the combination of Harris-Laplace and dense sampling should be used.

From the results shown in figure 5, it is observed that the SIFT variants perform substantially better than color moments, moment invariants and color histograms. The moments and histograms are not very distinctive when compared to SIFT-based features: they contain too little relevant information to be competitive with SIFT.

Because figure 5 shows only minor differences between SIFT and the four best color SIFT features (OpponentSIFT, WSIFT, rgSIFT and transformed color SIFT), the results per concept were analyzed. For bird, horse, motorbike, person and potted plant, it was observed that the features which perform best have scale-invariance and shift-invariance for light intensity (WSIFT and rgSIFT). The performance of the OpponentSIFT feature, which lacks scale-invariance compared to WSIFT, yields that scale-invariance, i.e. invariance to light intensity changes is important for these concepts. Transformed color SIFT includes additional invariance against light color changes and shifts when compared to WSIFT and rgSIFT. However, this additional invariance can make the feature less discriminative, because a reduction in performance is observed for some concepts. Overall, this is offset by a gain for other concepts. In fact, the lack of scale-invariance for light intensity of OpponentSIFT can be a strong point instead of a weak point: the intensity information in the feature potentially distinguishes concepts from false positives.

5.3 Experiment 3: Comparing color features on video benchmark

From the results shown in figure 6, the same overall pattern as for the image benchmark is observed: SIFT and color SIFT variants perform substantially better than the other color features. However, for this dataset, one of the color SIFT variants stands out: OpponentSIFT. An analysis on the individual concepts shows that

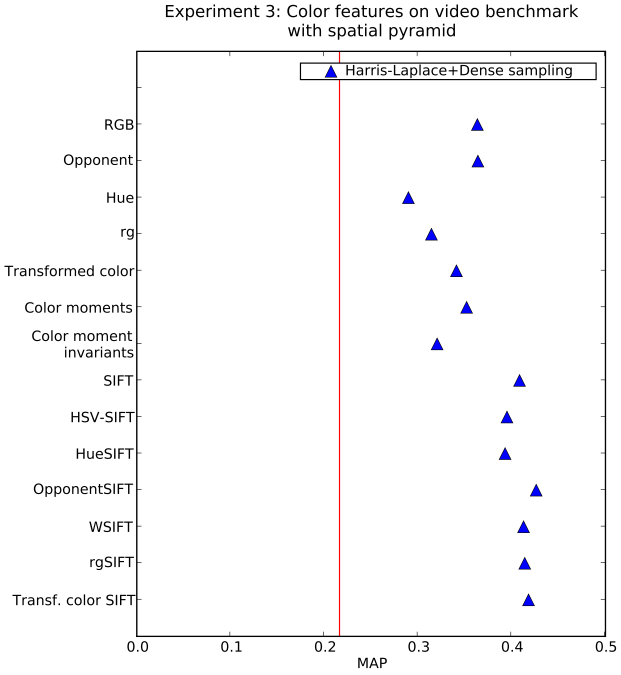

Figure 6: Evaluation of color features on a video benchmark, the Mediamill Challenge, averaged over 101 concepts from figure 3. Performance of the features shown are obtained using the combination of the Harris-Laplace detector and dense sampling, both with the spatial pyramid. The baseline performance provided by the Mediamill Challenge is indicated by a red line.

the OpponentSIFT feature performs best for building, outdoor, sky, studio, walking/running and weather news, *etc*. All these concepts either occur indoor or outdoor, but not both. Therefore, the intensity information present in the OpponentSIFT is very distinctive for these concepts. For the other color SIFT variants, there is a small performance gain for some concepts, for others there is a small loss.

5.4 Discussion

From the results of our experiments, it can be noticed that invariance to light intensity changes is concept-specific. For the video dataset, which consists of news broadcast videos, the light sources used are very diverse. Every television studio has its own lighting arrangement, all indoor scenes have different lighting because no flash is used when filming, *etc*. Therefore, in this setting the light intensity information can be highly discriminative for specific concepts which occur in a limited number of lighting conditions. However, there is also the analogous case for specific concepts which occur in widely varying lighting conditions. For these concepts, the color feature used should be invariant to light intensity and light color changes.

Because almost all color features are shift-invariant, the effect of light intensity variations on the performance cannot be observed easily. The color features which are sensitive to light intensity shifts are the three color histograms. Given that SIFT and its color variants show best performance, it can be derived that shift-invariance has no adverse effects on performance.

From the results, no firm conclusions can be drawn with respect to invariance to light color changes and shifts. Small performance gains are observed on a per-concept basis, but for other concepts there is a small loss. Overall, this does not make these color features stand out.

To illustrate that per-concept invariance is a viable strategy, we have performed a simple fusion experiment with the likelihood scores of SIFT and the best four color SIFT variants. These features all had similar overall performance on the PASCAL VOC dataset (MAP≈0.50). Combining these likelihood scores using product fusion [11], gives a MAP of 0.56. This convincing gain, with a naive method, suggests that the color features are complementary. Otherwise, overall performance would not have improved significantly. This is to be expected, as substantial differences on a per-concept basis were observed in section 5.2. Further gains should be possible, if the features with the right amount of invariance are fused, preferably using an automatic selection strategy.

For comparison, the best entry in the PASCAL VOC Challenge 2007, by Marszałek [17], has achieved a MAP of 0.59 using SIFT and HueSIFT features, additional Laplacian point sampling, extra image regions for the spatial pyramid and an advanced fusion scheme. The same simple fusion experiment performed on the Mediamill Challenge, where the input features have a MAP≈0.41, gives a score of 0.46. When compared to the baseline provided by the Mediamill Challenge (MAP=0.22), this is an improvement of 110%.

In summary, the level of invariance needed for color features is concept-specific. Results from a simple fusion experiment are very close to the state-of-the-art in the PASCAL VOC Challenge and exceed the baseline of the Mediamill Challenge by 110%.

6. CONCLUSION

In this paper, the distinctiveness of color features is assessed experimentally using two benchmarks from the image domain and the video domain, the PASCAL VOC Challenge 2007 and the Mediamill Challenge. From the results, it can be derived that invariance to light intensity changes and light color changes affects concept classification. The results show further that, the usefulness of invariance is concept-specific: for certain concepts, the lighting information itself is discriminative, whereas for other concepts invariance is needed.

Acknowledgements

This work was supported by the EC-FP6 VIDI-Video project.

7. REFERENCES

[1] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 401–408, Amsterdam, The Netherlands, 2007.

[2] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[3] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo. Large-Scale Multimodal Semantic Concept Detection for Consumer Video. In *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 255–264, Augsburg, Germany, 2007.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. `http://www.pascal-network.org/challenges/VOC/voc2007/`.

[5] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 524–531, San Diego, USA, 2005.

[6] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–271, 2003.

[7] J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, 2001.

[8] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[9] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 494–501, Amsterdam, The Netherlands, 2007.

[10] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *IEEE International Conference on Computer Vision*, pages 604–610, Beijing, China, 2005.

[11] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, New York, USA, 2006.

[13] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *British Machine Vision Conference*, pages 759–768, Norwich, UK, 2003.

[14] T. K. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, 2001.

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[16] R. Marée, P. Geurts, J. Piater, and L. Wehenkel. Random subwindows for robust image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 34–40, San Diego, USA, 2005.

[17] M. Marszałek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition, 2007. Visual Recognition Challenge workshop, in conjunction with IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil.

[18] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[19] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1-2):43–72, 2005.

[20] F. Mindru, T. Tuytelaars, L. Van Gool, and T. Moons. Moment invariants for recognition under changing viewpoint and illumination. *Computer Vision and Image Understanding*, 94(1-3):3–27, 2004.

[21] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[22] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *IEEE European Conference on Computer Vision*, volume 4, pages 490–503, Graz, Austria, 2006.

[23] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, pages 1470–1477, Nice, France, 2003.

[24] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 321–330, Santa Barbara, USA, 2006.

[25] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the ACM International Conference on Multimedia*, pages 421–430, Santa Barbara, USA, 2006.

[26] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluation of color descriptors for object and scene recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, 2008.

[27] J. van de Weijer, T. Gevers, and A. Bagdanov. Boosting color saliency in image feature detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):150–156, 2006.

[28] D. Wang, X. Liu, L. Luo, J. Li, and B. Zhang. Video diver: generic video indexing with diverse features. In *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 61–70, Augsburg, Germany, 2007.

[29] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia university's baseline detectors for 374 lscom semantic visual concepts. Technical Report 222-2006-8, Columbia University, 2007.

[30] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.