Robust Scene Categorization by Learning Image Statistics in Context

Jan C. van Gemert Jan-Mark Geusebroek Cor J. Veenman Cees G.M. Snoek
Arnold W.M. Smeulders

Intelligent Systems Lab Amsterdam,
Informatics Institute, University of Amsterdam,
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands.
jvgemert@science.uva.nl

Abstract

We present a generic and robust approach for scene categorization. A complex scene is described by proto-concepts like vegetation, water, fire, sky etc. These proto-concepts are represented by low level features, where we use natural images statistics to compactly represent color invariant texture information by a Weibull distribution. We introduce the notion of contextures which preserve the context of textures in a visual scene with an occurrence histogram (context) of similarities to proto-concept descriptors (texture). In contrast to a codebook approach, we use the similarity to all vocabulary elements to generalize beyond the code words. Visual descriptors are attained by combining different types of contexts with different texture parameters. The visual scene descriptors are generalized to visual categories by training a support vector machine. We evaluate our approach on 3 different datasets: 1) 50 categories for the TRECVID video dataset; 2) the Caltech 101-object images; 3) 89 categories being the intersection of the Corel photo stock with the Art Explosion photo stock. Results show that our approach is robust over different datasets, while maintaining competitive performance.

1. Introduction

Often, real world images only make sense when captured in context. For example consider an image of a harbor, a city skyline, or a conference meeting. Such scenes are captured more by the ensemble of objects, rather than by individual objects. Therefore, scene recognition differs from object recognition [3, 5, 10, 12] in that not only the foreground is the focus of recognition. Object recognition concentrates on the important task of detecting features relevant to one instance of an object, preventing as much as possible the inclusion of background features. Here we address the problem of scene categorization, *including* background and surrounding objects, that is, the context. Hence, we aim to contribute to content based image and video analysis by establishing a robust method for the learning and subsequent classification of scene categories.

Instead of using image features directly for scene categorization [20], several approaches [4, 11, 14, 15, 17, 18, 19, 21] make use of an intermediate image description step. This intermediate step consists of labeling a part of the image by its best representative out of a predefined codebook vocabulary. Using a codebook allows for density estimation [19], latent class analysis [4, 15, 17], and low level semantic grouping [11, 14, 18, 21]. An inherent problem of the codebook approach is choosing the vocabulary. If the vocabulary is too large, each part of the image will match to a single, unique, vocabulary element, which defies the purpose of a codebook. On the other hand, if the vocabulary is too small, several different image parts will be represented by the same vocabulary element. Thus, the codebook vocabulary determines the expressiveness and the discriminatory power of the method. In contrast, we propose to use the similarity to all codebook vocabulary elements, retaining expressiveness and discriminatory power.

In this paper, we exploit the statistical information locally available in images to categorize the scene. As shown by Torralba and Oliva [14, 18], scene categorization has strong correlation with the statistical structure within the image. Here, we provide a method for scene categorization, which does not need the input to be centered and oriented in a similar direction. Furthermore, the proposed method is robust over different datasets. We used one set of annotated video sequences to model video categories, object images, and photo stock collections. Moreover, the experiments are conducted with at least 50 categories using over ten thousand images. To our knowledge, this is the largest experimental evaluation, in number of categories and number of images, present in the literature.

The outline of the paper is as follows. The next section will give an overview of the visual features which effectively capture local image statistics. Subsequently, section 3 shows our method for learning context from local image statistics by learning the similarities of proto-concepts within images. Section 4 experimentally demonstrates our method on 3 dataset: 1) We show categorization results for 50 categories recognized on a large video collection of 160 hours of video. 2) To show the generality of our approach, we provide a comparison with state-of-the-art by learning and recognizing the 101 object categories in the Caltech collection. 3) to show the robustness of our approach, we will use the proto-concepts extracted from the video data to learn 89 categories from the Corel photo collection (16,500 images), and recognizing the learned categories in a completely different photo stock (ArtExplosion, 62,000 images). Finally, section 5 concludes the paper.

2. Visual Features

Modeling visual data heavily relies on qualitative features. Good features describe the relevant information in an image while reducing the amount of data representing the image. To achieve this goal, we use Weibull-based features [6]. By using Weibull-based features, we combine color invariance with natural image statistics resulting in an effective but compact description of local image content. Color invariance aims to remove accidental lighting conditions, while natural image statistics efficiently represent image data.

2.1. Natural Image Statistics

The statistical content of the scene provides robust cues for scene recognition [14, 18]. Hence, there is a direct relation between scene structure, and image statistics. In this paper, we exploit the statistical information locally available in images to categorize the scene. An example of such a categorization may be "close-up, indoor, outdoor, panorama". At a higher level of semantics, one may aim at categorizing the sort of objects in the image: "anchorman, explosion, boats, rural, city view, traffic jam". As will be demonstrated in this paper, both categorizations have strong correlations with the statistical structure of the scene.

We capture the local statistics of the image by applying Weibull-based features [6] where natural image statistics is used to effectively model texture information. For sake of completeness, we provide a short overview of Weibull-based features.

Texture is described by the distribution of edges for a certain region in an image. Hence, a histogram of a Gaussian derivative filter is used to represent the edge statistics. Since there are more non-edge pixels then there are edge pixels, a histogram of edge responses for natural images always has a peak around zero, i.e.: many pixels have no edge responses. Additionally, the shape of the tails of the distribution is often in-between a power-law and a Gaussian distribution. The tail emphasizes the long-range correlation between edge pixels in the image. A heavy power-law tail indicates a strongly contrasting object-background edge, whereas a Gaussian tail indicates a noisy, high-frequency texture region. The complete range of image statistics in natural textures can be well modeled with an integrated Weibull distribution [6]. This distribution is given by

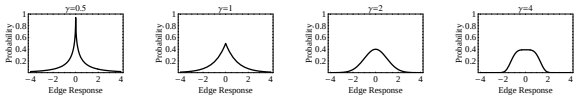

Figure 1. Some examples of the integrated Weibull distribution for $\beta = 1$, $\mu = 0$, varying values for $\gamma \in \{\frac{1}{2}, 1, 2, 4\}$.

$$\frac{\gamma}{2\gamma^{\frac{1}{\gamma}}\beta\Gamma(\frac{1}{\gamma})} \exp\left\{-\frac{1}{\gamma}\left|\frac{r-\mu}{\beta}\right|^{\gamma}\right\} \quad , \qquad (1)$$

where r is the edge response to the Gaussian derivative filter and $\Gamma(\cdot)$ is the complete Gamma function, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-1}dt$. The parameter β denotes the width of the distribution, the parameter γ represents the peakness of the distribution, and the parameter μ denotes the origin of the distribution. See figure 1 for examples of the integrated Weibull distribution.

The integrated Weibull distribution can be estimated from a histogram of filter responses with a maximum likelihood estimator (MLE). The parameters μ, β and γ are estimated by taking the derivatives of the integrated Weibull distribution to the respective parameters and setting them to zero. The parameters β and γ are dependant on each other, therefore a binary search scheme is utilized to estimate the best β and γ combination.

Since the integrated Weibull distribution characterizes edge responses, the parameters of the distribution correspond to different image properties. The β parameter represents the width of the distribution. A high value of β corresponds to a wide distribution which indicates an image with high contrast. The γ parameter denotes the slope of the distribution. A low value of γ (< 1) represents a highly peaked distribution, which corresponds to an image with smooth surfaces. A medium value of γ $(1 < \gamma < 2)$ indicates a smooth distribution, which represents Gaussian noise-like images. A high value of γ (> 2) is an indicator of a histogram that does not follow a Weibull distribution. Specifically, an image with a regular pattern, for example the beams of the American flag, produces a histogram that has multiple peaks. The MLE estimator of the integrated Weibull distribution will represent multiple peaks in the histogram by a smooth and flat distribution, represented

by a high value of γ. The μ parameter represents the mode of the distribution. The position of the mode is influenced by uneven illumination and colored illumination. Hence, to achieve color constancy the values for μ may be ignored.

To assess the similarity between two integrated Weibull distributions, a goodness-of-fit test is utilized. The measure is based on the integrated squared error between the two cumulative distributions, which is obtained by the Cramér-von Mises statistic,

$$C^2 = \int_0^1 \left[F(x) - G(x)\right]^2 dF(x) \quad , \tag{2}$$

where F is the test distribution, and G represents the target distribution, where both are cumulative distributions. For two Weibull distributions with parameters β_F, γ_F and β_G, γ_G a first order Taylor approximation yields the log difference between the parameters. Therefore, we define a measure of similarity between two Weibull distributions is given by the ratio of the parameters,

$$C^2(F,G) = \sqrt{\frac{\min(\beta_F,\beta_G)}{\max(\beta_F,\beta_G)}\frac{\min(\gamma_F,\gamma_G)}{\max(\gamma_F,\gamma_G)}} \quad . \tag{3}$$

In summary, Weibull-based features provide a texture descriptor based on edges. Moreover, the features rely heavily on natural image statistics to compactly represent the visual information. For a more detailed elaboration on Weibull-based features, see [6].

2.2. Color Invariant Edge Detection

Here we combine color invariant edge responses with natural image statistics to end up with color invariant Weibull-based features. Color invariance aims to remove accidental lighting conditions, while Weibull-based features efficiently represent image statistics.

We first decorrelate the RGB channels by a linear transformation to an opponent color representation. Advantage of the use of an opponent color space is that color values are decorrelated. Hence, for a distinctive image content descriptor, we may as well use the marginal, one-dimensional, distributions for each of the color channels. This in contrast to the histogram of the full 2D chromatic or 3D color space (see e.g. [2, 8]).

Further decorrelation of color information can be achieved by using photometric invariant edge detectors. The invariant W (notation from [7])) measures all intensity fluctuations except for overall intensity level. That is, edges due to shading, cast shadow, and albedo changes of the object surface. These invariants are equivalent to Gaussian derivative filters for color images, where 6 orthogonal derivatives may be distinguished. W_x, W_y detect edges in intensity, whereas $W_{\lambda x}, W_{\lambda y}$ and $W_{\lambda\lambda x}, W_{\lambda\lambda y}$ detect edges in the two orthogonal chromatic color components.

Thus, color invariant edge responses, are invariant to changes in intensity, and decorrelate the RGB channels, allowing the weibulls to be computed on marginal densities.

3. Contextures: Regional Texture Descriptors and their Context

Building towards semantic access to image collections, we aim to decompose complex scenes in proto-concepts like vegetation, water, fire, sky etc. These proto-concepts provide a first step to automatic access to image content [21]. Given a fixed vocabulary of proto-concepts, we assign a similarity score to all proto-concepts for all regions in an image. Different combinations of a similarity histogram of proto-concepts provide a sufficient characterization of a complex scene. We introduce the notion of contextures, where global texture and local texture information and their context are used to describe visual scene information.

By using the similarity to all vocabulary elements, we introduce an alternative to codebook approaches [4, 15, 17, 19, 21]. A codebook approach uses the single, best matching vocabulary element to represent an image patch. For example, given a blue area, the codebook approach must choose between water and sky, leaving no room for uncertainty. We propose to use the distances to all vocabulary elements. Hence, we model the uncertainty of assigning an image patch to each vocabulary elements. By using similarities to the whole vocabulary, our approach is able to model scenes that consist of elements not in the codebook vocabulary.

3.1. Region Annotation of Proto-Concepts

In order to recognize concepts based on low-level visual analysis, we annotated 15 different proto-concepts: building (321), car (192), charts (52), crowd (270), desert (82), fire (67), US-flag (98), maps (44), mountain (41), road (143), sky (291), smoke (64), snow (24), vegetation (242), water (108), where the number in brackets indicates the number of annotation samples of that concept. These proto-concepts are chosen by their relevance for concept detection in the TRECVID video benchmark. Although they seem to be tuned to the problem at hand, we will show these concepts to generalize (including the annotation effort) to various datasets. Fig. 2 shows an example of some regional annotations. We use the TRECVID 2005 [13] common annotation effort as a basis for selecting relevant shots containing the proto-concepts. In those shots, we annotated rectangular regions where the proto-concept is visible for at least 20 frames.

For each of the proto concepts, visual characteristics are captured by their Weibull-based features as described above.

Sky Building Road

Figure 2. Three examples of annotated regions in video.

Figure 3. An example of dividing an image up in overlapping regions. Here, the region size is a $\frac{1}{2}$ of the image size for both the x- and y-dimension. The regions are uniformly sampled across the image with a step size of half a region. Sampling in this manner identifies nine overlapping regions.

3.2. Region descriptors

The visual detectors aim to decompose an image in similarities to proto-concepts like vegetation, water, fire, sky etc. To achieve this goal, an image is divided up in several overlapping rectangular regions. The regions are uniformly sampled across the image, with a step size of half a region, see figure 3 for an example. The region size has to be large enough to assess statistical relevance, and small enough to capture local textures in an image. We utilize a multi-scale approach, using small and large regions.

A visual scene is characterized by both global as well as local information. For example, a picture with an aircraft in mid air might be described as "sky, with a hole in it", sky being globally present in the image except for a local distortion: the aircraft. To model this type of information, we use a proto-concept occurrence histogram where each bin is a proto-concept. The values in the histogram are the similarity responses of each proto-concept, to the regions in the image.

We use the proto-concept occurrence histogram to characterize both global and local texture information. Global information is described by computing an occurrence histogram accumulated over all regions in the image. Local information is taken into account by constructing another occurrence histogram for only the response of the best matching region. For each proto-concept, or bin, b the accumulated occurrence histogram and the best occurrence histogram are constructed by,

$$H_{accu}(b) = \sum_{r \in R(im)} \sum_{a \in A(b)} C^2(a,r) \quad , \qquad (4)$$

$$H_{best}(b) = \arg\max_{r \in R(im)} \sum_{a \in A(b)} C^2(a,r) \quad , \qquad (5)$$

where $R(im)$ denotes the set of regions in image im, A(b) represents the set of stored annotations for proto-concept b, and C^2 is the Cramér-von Mises statistic as introduced in equation 2. We denote a proto-concept occurrence histogram of an image as a contexture for that image. We have chosen this name, as our method incorporates texture features in a context. The texture features are given by the use of Weibull-based features, using color invariance and natural image statistics. Furthermore, context is taken into account by the combination of both local and global region combinations.

The contexture H_{accu} counts the relative amount of proto-concepts present in a scene, hence *how much* of a proto-concept is present in a scene. The contexture H_{accu} is important in characterizing, for example, airplanes and boats. In these cases, the accumulated histogram indicates the presence of a large water body or a large area of sky. The contexture H_{best} only indicates the presence of proto-concepts, hence indicates *which* proto-concepts are present in a scene. In this way, constellations of proto-concept indicate scene type without specifying the relative area each proto-concept should occupy. This is of importance in characterizing, for example, military actions in the middle east, where the combined presence of road, desert, and fire, turns out to be very effective. Note that, by using occurrence histograms and dense sampling over the image, the proposed method is translation invariant, thus, the exact layout of the scene is not strictly enforced. Opposed to [14], placing objects in the centre of the scene, and strictly aligning them in a similar direction is not necessary for our categorization scheme.

In contrast to codebook approaches, our method is not limited to the visual categories that can be described by the vocabulary of proto-concepts. Not every image contains proto-concepts like 'sky', 'vegetation', 'water'. Scenes where the specific proto-concepts do not occur can nevertheless be described by contextures. This is the case, since the similarity to a proto-concept is used, not the proto-concept itself. A robust and consistent similarity measure will give similar values for similar scenes. For scenes that belong to the same visual category, there is some common visual denominator that ties the scenes to the category. Hence, there will be a correlation between the contextures of scenes that belong to the same category. For example, an office scene might consist of large surfaces with sharp edges (desks) and multicolored highly textured and oriented regions (books). The similarity to proto-concepts

like 'sky' and 'vegetation' will not be high since none of the proto-concepts are present. However, the responses of the proto-concepts will be the same for another office scene, because this new scene will consist of similar regions. Thus, a scene can be expressed in a degree of similarity to a vocabulary of proto-concepts, without containing any of the proto-concepts.

Learning of scene categories is approached by default machine learning techniques. The contextures are extracted from example images, human labeled to belong to a given category, and subsequently fed into a support vector machine (SVM) with a radial basis function for scene category learning.

4. Experiments

Contextures can be computed for different parameter settings. Specifically, we calculate the contextures at scales $\sigma = 1$ and $\sigma = 3$ of the Gaussian filter. Furthermore, we use two different region sizes, with ratios of $\frac{1}{2}$ and $\frac{1}{6}$ of the x-dimension and y-dimensions of the image. The combination of all these parameters yields a single vector, which is used for scene classification.

4.1. TRECVID video benchmark

The TRECVID video benchmark 2005 [13] provides nearly 170 hours of news video (English: CNN, NBC, MSNBC; Chinese: CCTV4, NTDTV; Arabic: LBC). The goal is to retrieve shots from this collection, which are relevant to a predefined topic. The National Institute of Standards and Technology (NIST) provides the video collection to all participants, and scores the returned rankings by human evaluation.

Video retrieval is evaluated by the relevance of a shot, while contextures are based on one image. To generalize our approach to shot level, we extract 1 frame per second out of the video, and then aggregate the frames that belong to the same shot. We use two ways to aggregate frames: 1) average the contexture responses for all extracted frames in a shot and 2) keep the maximum response of all frames in a shot. This aggregation strategy accounts for information about the whole shots, and information about accidental frames, which might occur with high camera motion. However, since we do not use keyframes, we lose information about exactly identical shots, like commercials.

We learned 50 categories on the video data, shown in figure 4. The results provided here gives an impression of the quality of visual only detection by using our method of scene categorization, compared to state-of-the-art video retrieval. For all 50 visual concepts we extracted from the video, 10 categories are evaluated by NIST. In figure 5 we provide the average precision for these 10 categories for our method against the best and the median result for all 33

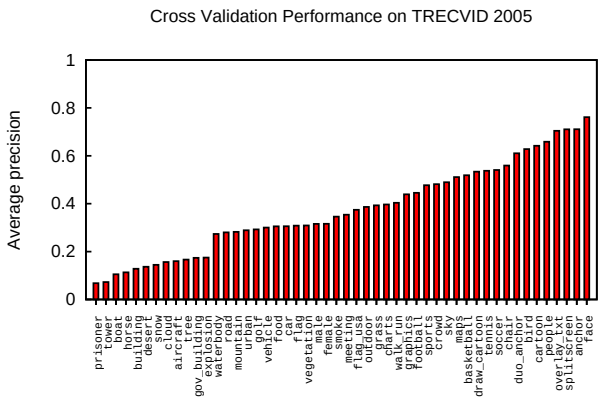

Figure 4. Performance measured in average precision of 50 visual-only detectors on TRECVID data. The score was computed by three fold cross-validation.

other participants. We do not get the best results however obtain competitive results to all participants.

Overall, the proposed scene categorization turns out to work effectively for 1) scenes where spatial context is uniform, like individual sports (soccer, tennis, basketball, football), 2) typical studio settings (anchor, face, spitscreen), and 3) well constrained environments (e.g., "chairs" and "tables" coincides with interview settings or political items in news). Performance for combinations of these categories are not well learned from examples alone (see e.g. sports), and need a higher level aggregation step. Furthermore, natural scene categories are well represented by the proposed scheme, for example mountains, waterbody, vegetation, smoke. Visual inspection shows that the scene categorization is well able to generalize learned concepts to an unseen test set. Note that for the 10 evaluated concepts, TRECVID results for at least 3 concepts (waterbody, cars, mountains) are dominated by commercials (identical copy detection), for which we did not make an additional effort.

4.2. Caltech 101 object Categories

In the previous section we gave an impression of our scene categorization on a large collection of video data. From the training set of the TREC video collection, the proto-concept annotations have been extracted. Hence, the proto-concepts are tuned to the type of data (compression, quality), and possibly include domain specific information. An important research question is if the learned similarity histograms of proto-concepts, at the heart of our method, easy generalize to other domains and image qualities. Here, we compare performance on a standard collection of web-images: the Caltech 101 object categories.

In figure 6, we compare classification performance

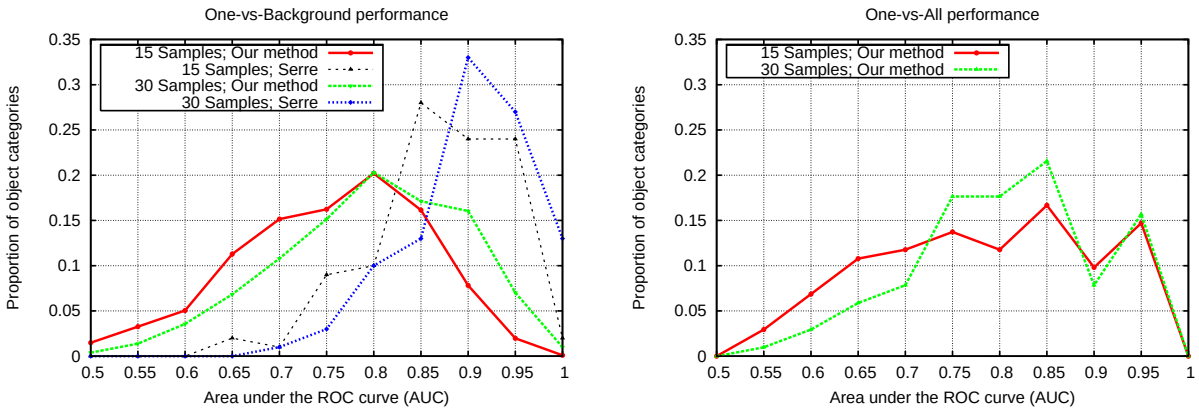

Figure 6. Performance histogram on the Caltech 101 object dataset, for different numbers of training examples: (left) one class vs. background class and (right) one class versus all other classes.

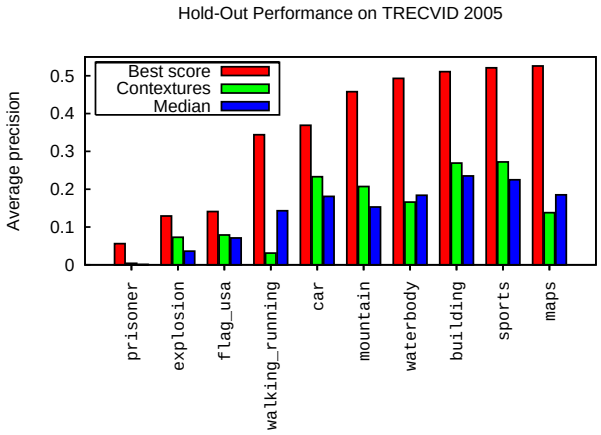

Figure 5. Best, median and our average precision scores for the 10 concepts on hold-out data, evaluated by NIST.

against Serre et al. [16]. For recognition of each single categories against a background class of Google images, (one vs background), performance is not as good as by Serre et al. This can be explained by the fact that the Caltech collection contains several manipulation artifacts, in that objects have been centered and orientation has been normalized within each category. Furthermore, several computer graphics and cartoons are included in object categories, and, more important, convey a large portion of the background class. Hence, our natural image statistics based description is not too adequate here. However, a perfect classification of foreground-background gives no indication of performance if an unknown scene has to be classified. You can have perfect one-vs-background, and at the same time being poor in separating all 101 classes. We expect our method to perform better under one-against-all since in that case the number of cartoon images becomes less dominant.

Our performance on multiclass classification for 15 training samples is 33.2% correct classification, and for 30 training samples 42.3% correct classification (chance 1/101 is below 1%). Compared to the paper by Fei Fei et al. [3], who reach 16% correct classification for 15 examples. Serre et al. [16] has 35% for 15 examples, and 42% for 30 examples, comparable to our results. Holub et al. [9] obtain 40% classification accuracy for 20 training examples. Berg et al. [1] reach best performance of 45% with 15 examples.

Note that we obtain a similar performance as the methods cited above, with a limited feature set derived from only 12 Gaussian derivative filters. Hence, our method generalizes generalizes beyond the original domain of video to web images.

4.3. Corel vs. ArtExplosion

To further evaluate the robustness of our approach, we applied scene categorization on a photo stock. In this experiment, we investigate if scene categories learned from one collection can be applied to a different collection. Note that, from a machine learning perspective, this is a more challenging task then obtaining a training and test set by subdividing a homogeneous collection. We use the Corel and ArtExplosion commercially available photo stock, and take the intersection in categories between the two as dataset, see figure 7. Hence, we have 89 categories, on one side 16,499 Corel images, ranging from 99 to 700 examples per category; on the other side 62,072 ArtExplosion images ranging from 26 to 4,896 examples per category.

We learned the categories for the Corel collection and ArtExplosion collection separately, and applied the models learned from the one collection to retrieve the categories from the other collection, see figure 8 and figure 9. Main result is that geographical locations (countries, cities) are not performing well: for 43 location concepts, there are 39 performing below 0.1 average precision. The remaining 4 locations are (average precision on cross validation between brackets): Italy (0.11); Yemen (0.12); Egypt (0.16);

Africa Agricltr (Architct Agricltr)(Agriculture) Alaska Architecture (Archit Archit Arct Euro Arct Urbn Arct Wrld)(Architecture) Arizona Austrlia Aviation (Aviation WarPlane)(Aircraft Air_Show) Ballet Balloon (Balloon1 Balloon2)(Balloons) Boats (Boats Nautical TallShip)(Nautical Boats_&_Ships) Canada Caribean Car (Car_Brit Car_Old1 Car_Old2 Car_Perf Car_Race Car_Rare Car_BGs)(Auto_Racing Automobiles Classic_Autos_&_Motorcycles) Castle (Castle Castle Castle Castles)(Castles) China (China China II)(China) Religion (Churches Relgnl Relgnll Relgnlll Worship)(Religion Synagogues_&_Churches) Colorado Sci-Tech (Indstry Industry Sci&Tech ComTech)(Science_&_Industry Industrial) Work (Work)(People_at_Work) CstaRica (CstaRica)(Costa_Rica_&_Guatemala) Cuisine (Cuisine Cuisines Food Food1 Food2)(Food) CzechRep (CzechRep)(Czechoslovakia) Sunsets (Sunsets Dawn&Dsk Skies Skies II)(Sunset_&_Sky Clouds_&_Sky Sunrise_&_Sunset) Desert (Desert Desert II)(Desert) Egypt (Egypt1 Egypt2)(Egypt Egyptian_Art_&_Architecture) England Garden (Estate Plants Gardens Flwr Beds)(Gardens) Europe (Europe Europe_E Europe_S Euro_Scn)(Europe) Evrglade Fireworks (Firewrk1 FireWrk2)(Fireworks) Flags Florida Flower (Flower1 Flower2 Flower Flower II)(Plants Flowers) Forest (Forest Forests)(Trees_&_Plants Trees) France Golf Greece Hawii (Hawii)(Hawaii Hawaii) HongKong Horses (Horses Rodeo)(Horse_Sports) India (India_I India_II)(India) Interior (Interior)(Interiors) Italy Japan (Japan Japan1 Japan2)(Japan) Jewelry Kenya Korea Landmark (Landmark Landmrks)(Landmarks Cityscapes Real_Estate) Ruins (Ruins Ruins1 Ruins2 Ruins3)(Ancient_Ruins) LightHouse (LightHse)(Lighthouses) London MiddleEast (MdleEast)(Middle_East) Mexico (Mexico MexicoCt)(Mexico) Mountain (Mountain)(Mountain_Scenes) Music (Music)(Musical_Instruments) Namibia NwMexico (NwMexico)(New_Mexico) NwZealnd (NwZealnd)(New_Zealand) Oregon People (People1 People2)(People People_General) Quebec Signs (Rd_Signs)(Universal_Signs Signs) Recreation (Recreatn)(Recreation) Road (Road&Hwy)(Roads_of_the_World Highways_&_Byways) Rock-Gem (Rock&Gem)(Rocks_&_Gems) RockForm (RockForm)(Rockscapes) Rural Russia San_Fran (San_Fran)(San_Francisco) Scotland Sculpture (Sculpt Sculpt I Sculpt II)(Statuary_&_Objects) Seasons (Seasons)(The_Four_Seasons) SE_Asia (SE_Asia)(Asia) SF_Doors (SF_Doors)(Doors_&_Windows) Space (Space Space1 Space2)(Space Science) Sport (Sports Sports II)(Sports) SubSea (SubSea1 SubSea2 SubSea3 SubSea II)(SubSea_I Under_the_Sea) Thailand Train (Train1 Train2 Train3)(Trains_&_Tracks Trains) Transport (Trans1 Trans II)(Transportation) Utah Vietnam (Vietnam)(Vietnam_Voyage) Virginia Wash_DC (Wash_DC)(Washington_DC) Waterscape (Waterfal WaterScn Waterway Waves)(Water Waterscapes) WetSport (WetSport Windsurf)(Water_Sports) Winter (Winter)(Winterscapes) Yemen Zimbabwe

Figure 7. The 89 base concepts, with corresponding categories in Corel and Artexplosion. The main concept is followed by its constituent categories in brackets. The concepts in brackets are the corresponding categories in Corel and ArtExplosion, respectively.

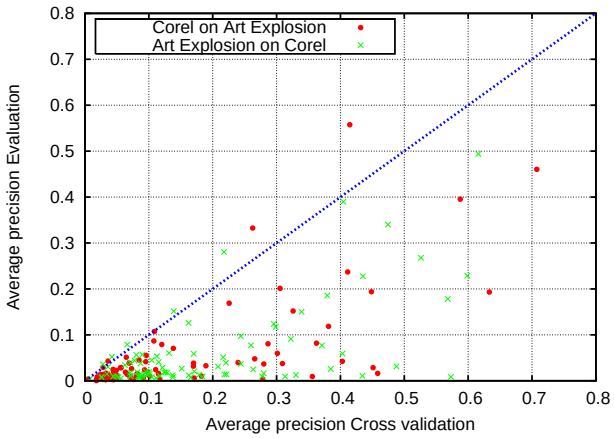

Figure 8. Performance measured in average precision on training and evaluation sets. The x-axis is the inter-set cross validation performance, where the y-axis displays the performance scored on the other set.

Utah (0.26). The relative high scores are explained by a large overlap in similar places photographed in both Corel and ArtExplosion. Hence, we draw the conclusion that geographical locations can only be categorized by learning and retrieving typical landmarks.

To evaluate categories which do perform well, we made a human judged ground truth of the top-100 results of the non-geographical locations categories. The results are given in figure 10, and some examples are shown in figure 11. Note that in these 46 categories, still 11,000 Corel images and 50,465 ArtExplosion images are available. For the top 100 results the Corel models evaluated on ArtExplosion score on average better than the ArtExplosion models

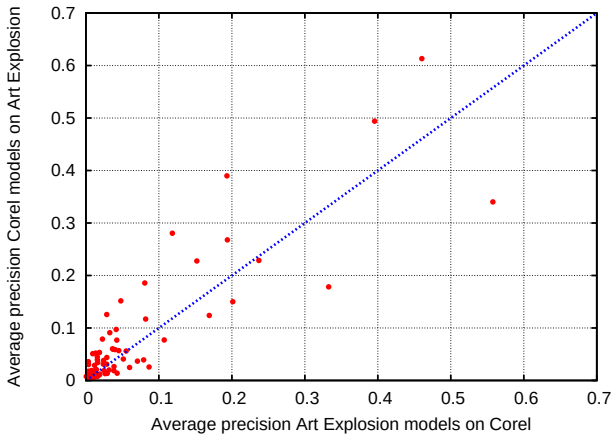

Figure 9. Performance of models evaluated on different sets.

evaluated on Corel. On average, the Corel on ArtExplosion measured with strict category membership has 21% correct, while manually counting the output of the methods shows it has 33% correct. Conversely, the ArtExplosion evaluated on Corel, with strict categories has 17% correct, and with manual counting of the output shows 24% correct. Categories that are consistently well performing are: architecture, people, wetsport, waterscape, mountain, subsea, flags, balloon, signs, boats, forest, aviation, fireworks, flower, and sunset. Note that building, water, flag, sky, vegetation are proto-concepts learned from the TRECVID video collection. These concepts appear to be transposed to the stock photo collection, increasing performance for related categories.

5. Conclusions

In this paper we have presented scene category classification by learning the occurrence of proto-concepts in images. We compactly represent these proto-concepts by using color invariance and natural image statistics properties. By exploiting similarity responses as opposed to strict selection of a codebook vocabulary, we have been able to generalize these proto-concepts to be applicable in general image collections. We have demonstrated the applicability of our approach in a) learning 50 scene categories from a large collection of news video data; b) a collection of 101 categories of web images; and c) two large collections of photo-stock images, comprising 89 categories, where categories are learned from one and categorized from the other.

In conclusion, we have provided an effective scheme for scene categorization. An important contribution is scalability, showing that the proposed scheme is effective in capturing visual characteristics for a large class of concepts, over a wide variety of image sets. Where specific methods may have better performance for specific datasets, we have shown a method which is neither tuned nor optimized

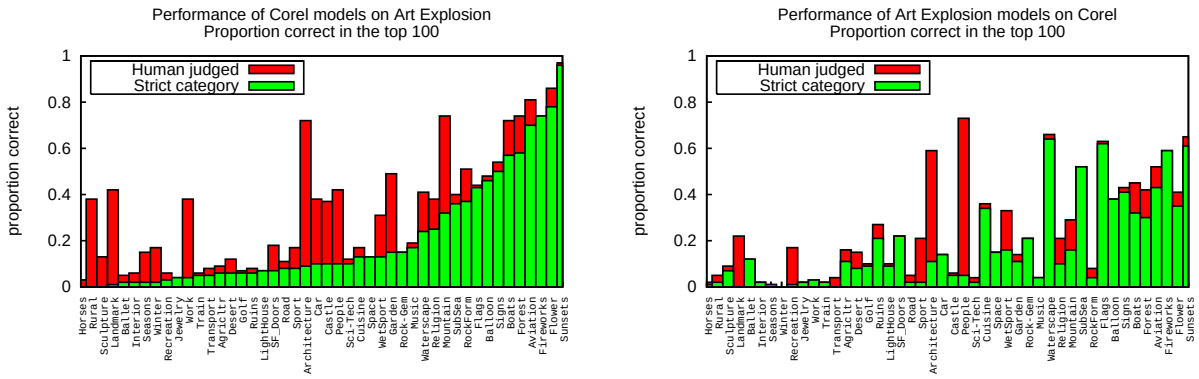

Figure 10. Percentage correct classification in the Top 100 results for 46 categories. The ground truth is contrasted with the given categories, where countries and cities are not included as a ground truth can not be established.

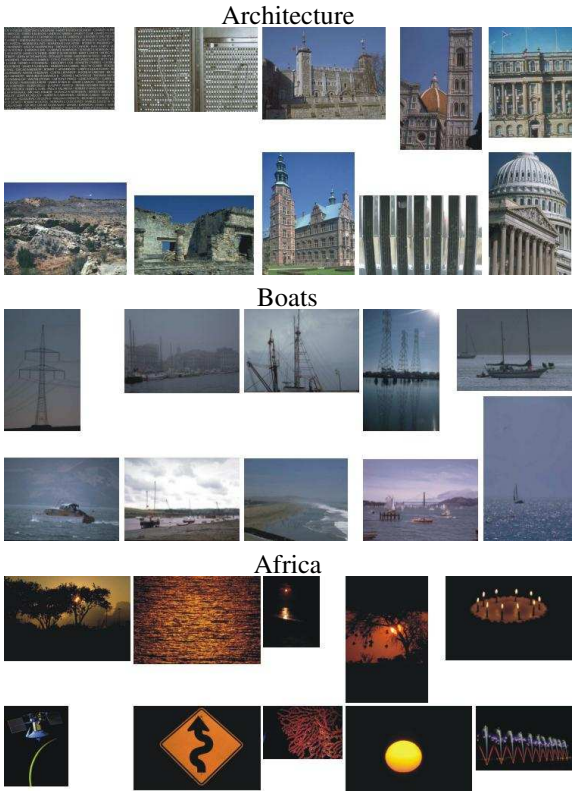

Figure 11. Examples of top 10 results. Only for Africa, according to the categories, none are correct.

in parameters for each collection, other than the TRECVID video dataset. Hence, the method has proven to robustly categorize scenes from learned context.

References

[1] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *CVPR*, pages 26–33, 2005. 6

[2] F. Ennesser and G. Medioni. Finding Waldo, or focus of attention using local color information. *PAMI*, 17:805–809, 1995. 3

[3] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *WGMBV*, 2004. 1, 6

[4] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005. 1, 3

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003. 1

[6] J. M. Geusebroek and A. W. M. Smeulders. A six-stimulus theory for stochastic texture. *IJCV*, 62(1/2):7–16, 2005. 2, 3

[7] J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts. Color invariance. *PAMI*, 23(12):1338–1350, 2001. 3

[8] T. Gevers and A. W. M. Smeulders. Color based object recognition. *Pat. Rec.*, 32:453–464, 1999. 3

[9] A. D. Holub, M. Welling, and P. Perona. Combining generative models and fisher kernels for object class recognition. In *ICCV*, 2005. 6

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004. 1

[11] J. Luo and M. R. Boutell. Natural scene classification using overcomplete ica. *Pat. Rec.*, 38(10):1507–1519, 2005. 1

[12] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *ICCV*, 2001. 1

[13] NIST. TRECVID Video Retrieval Evaluation, 2001–2005. http://www-nlpir.nist.gov/projects/trecvid/. 3, 5

[14] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 1, 2, 4

[15] P. Quelhas, F. Monay, J. M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *ICCV*, 2005. 1, 3

[16] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *CVPR*, pages 994–1000, 2005. 6

[17] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Describing visual scenes using transformed dirichlet processes. In *NIPS*, 2005. 1, 3

[18] A. Torralba. Contextual priming for object detection. *Int. J. Comput. Vision*, 53(2):169–191, 2003. 1, 2

[19] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117–130, 2001. 1, 3

[20] A. Vailaya, A. K. Jain, and H. Zhang. On image classification: city images vs. landscapes. *Pat. Rec.*, 31(12), 1998. 1

[21] J. Vogel and B. Schiele. Natural scene retrieval based on a semantic modeling step. In *ICVR*, Dublin, Ireland, July 2004. 1, 3