Appears in IEEE International Conference on Image Processing, San Diego, 2008.

A SYSTEMATIC STUDY OF THE ROLE OF CONTEXT ON IMAGE CLASSIFICATION

Nikhil Rasiwasia, Nuno Vasconcelos

University of California San Diego
Department of Electrical and Computer Engineering

ABSTRACT

We present the results of a systematic study of the *contextual gain hypothesis* for image classification. This hypothesis relates the traditional strategy of direct visual classification (DVC), and an alternative strategy based on indirect contextual classification (ICC). DVC is composed of classifiers that operate directly on pixel or feature based image representations. ICC relies on DVC to label images with respect to a pre-defined set of contextual semantic features. Image classification is then performed by a classifier that operates on the semantic space of these classifier outputs. The contextual gain hypothesis states that, in this semantic space, it is possible to design classifiers with better accuracy than those achievable with DVC. A framework for the systematic comparison of the DVC and ICC strategies is introduced, and an extensive comparison of the performance of the two strategies is carried out. Its results strongly suggest that the contextual gain hypothesis holds.

***Index Terms*—** Image analysis, image classification, contextual learning, semantic space, image retrieval.

1. INTRODUCTION

Image classification is an important problem for various areas of image processing, including image and video retrieval, texture analysis, and the design of recognition or surveillance systems. While the last decade has produced significant advances with respect to this problem, the fundamental strategy for classifying images has not changed significantly from what has been the norm for a number of decades. It consists of 1) identifying a number of visual classes of interest, 2) designing a set of appearance based features such as image pixels, edge responses, texture features etc, that are optimally discriminant for those classes, 3) postulating a model for the classification of those features, and 4) relying on sophisticated mathematical tools to fit that to examples. We refer to this strategy as *direct visual classification* (DVC), because the associated classifiers rely on image representations which are either direct visual appearance features or derived by simple deterministic mappings of those features.

While there is no question that DVC will retain a predominant role in the future of image understanding, it is not as clear that it will be *sufficient* to solve all classification problems. In fact, there is so far little evidence that it can solve all but a small class of problems (such as face detection) with accuracies comparable to those of biological vision. One striking property of the latter, at least in what concerns humans, is that it rarely seems to ground decisions exclusively on low-level visual features. This has been well documented in psychophysics, through unambiguous evidence that scene interpretation depends on *context* [1, 2]. By this, it is usually meant that detection of an object of interest (e.g. a locomotive) is facilitated by the presence, in the scene, of other objects (e.g. railroad tracks or trains) which may not themselves be of interest.

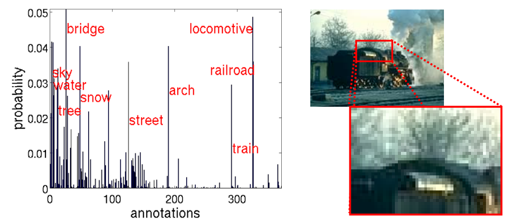


Fig. 1. An image and its associated SMN (see Sec. 2.3). Note that, while most of the concepts of largest probability are present in the image, the SMN assigns significant probability to "bridge" and "arch". This is due to the presence of a geometric structure similar to that of "bridge" and "arch", shown on the image close-up.

The presence of these *contextual cues* (e.g. that locomotives are usually on tracks and pull trains) increases the detection rate for the object of interest. This is illustrated in Figure 1, where we present the posterior probabilities of a locomotive image belonging to a number of visual concept classes, according to a number of direct visual detectors trained on those classes. Although, posterior probability of "bridge" is slightly higher than that of "locomotive", due to the presence of an 'arch-like structure in the locomotive's rooftop, a context-sensitive classifier could still assign the image to the "locomotive" class by noting that the contextual cues "railroad", and "train" also have high posterior probability. We refer to this classification strategy as *indirect contextual classification* (ICC), since classifiers operate on higher-level, *contextual cues* which provide additional information for the classification process.

Indirect contextual classification has been previously studied by a number of authors[3, 4, 5, 6]. In [3], Wolf et. al. presents a concise review of various techniques employed to integrate contextual cues in the classification architecture. Inspite of these advances, the fundamental questions of whether there is an intrinsic value to using ICC for image classification, remains poorly understood. Moreover, the complexity of learning and inference in existing algorithms, makes it impractical to *systematically* study relevant questions pertaining to ICC, for example the question of how classification performance depends on richness of the set of contextual cues.

In this work, we address the problem of whether there is a benefit in considering context for classification and present a *systematic* study of ICC. For this, we introduce a framework for objective comparison of the two - visual and contextual strategies. In particular, we adopt two image classification systems that, while simple, have been shown to perform well in image retrieval context. The first is a DVC system [7], which evaluates similarity in strict visual terms. The second is an ICC system [8], which evaluates similarity at the contextual level in two stages. First, in the *semantic labeling* stage, a bank of *parallel* and *independent* direct visual classifiers are trained for the detection of pre-specified semantic concepts. An image is thus represented as the posterior concept probabilities, which constitutes a *higher-level semantic space* in which all classification decisions are ultimately made. In the second stage, these posterior concept probabilities are fed to a contextual classifier, that returns the database images with closest concept posterior distribution to that of the query image. The two adopted systems are identical in all aspects of 1) visual representation, and 2) classification architecture, which makes the difference in classification strategy the only explanation for the differences in performance.

An extensive comparison of classification accuracy is performed on a diverse set of image databases. The results are very clearly in support of the hypothesis that there is a *contextual gain*. The dependence of this gain on a number of factors, including the *accuracy of the underlying DVC architecture*, and the *number of informative semantic dimensions*, are then systematically characterized. It is shown that the contextual gain increases with these two factors.

2. PROPOSED FRAMEWORK

In this section, we first introduce the image representation used at the visual and the contextual level, and then describe the DVC and ICC architecture, compatible with the minimum probability of error (MPE) classification [7].

2.1. Image Representation

The starting point for all image classification systems is an image database $\mathcal{D} = \{\mathcal{I}_1, \ldots, \mathcal{I}_D\}$. At the *visual-level*, images are observations from a random variable $\mathbf{X}$, defined on some visual feature space $\mathcal{X}$. Each image is considered an observation from a class, determined by a random variable Y. An image is represented as a set of *n feature vectors* $\mathcal{I} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}, \mathbf{x}_i \in \mathcal{X}$. Although any type of visual features are acceptable, we only consider *localized features*, i.e., features of limited spatial support, assumed to be sampled independently. An image is thus represented as,

$$P_{\mathbf{X}|Y}(\mathcal{I}|y) = \prod_j P_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}_j|y). \tag{1}$$

and a density estimation [9] procedure is used to estimate the distributions $P_{\mathbf{X}|Y}(\mathbf{x}|y)$.

At the semantic-level, the database $\mathcal{D}$ is augmented with a vocabulary $\mathcal{L} = \{w_1, \ldots, w_L\}$ of semantic concepts or keywords w_i, and each image I_i with a caption $\mathbf{c}_i$, making $\mathcal{D} = \{(\mathcal{I}_1, \mathbf{c}_1), \ldots, (\mathcal{I}_D, \mathbf{c}_D)\}$. Note that $\mathbf{c}_i$ is a binary L-dimensional vector such that $\mathbf{c}_{i,j} = 1$ if the i^{th} image was annotated with the j^{th} keyword in $\mathcal{L}$. Concepts are drawn from a random variable W, which takes values in $\{1, \ldots, L\}$, so that $W = i$ if and only if $\mathbf{x}$ is a sample from the concept w_i. Each concept induces a probability density $\{P_{\mathbf{X}|W}(\mathbf{x}|i)\}_{i=1}^L$ on $\mathcal{X}$, from which feature vectors are drawn. Images are assumed to be independently sampled from concept distributions

$$P_{\mathbf{X}|W}(\mathcal{I}|w) = \prod_j P_{\mathbf{X}|\mathbf{W}}(\mathbf{x}_j|w), \tag{2}$$

and the density estimation procedure used in (1) is also used to estimate the distributions $P_{\mathbf{X}|W}(\mathbf{x}|w)$.

2.2. Direct Visual Classification System

The DVC system operates at the visual level. In the absence of class labels, each image is considered an observation from a different class, i.e the random variable Y is then defined on $\{1, \ldots, D\}$. Given a query image $\mathcal{I}_q$, the MPE decision rule is to assign it to the class of largest posterior probability, i.e.

$$y^* = \arg\max_y P_{Y|\mathbf{X}}(y|\mathcal{I}_q). \tag{3}$$

We refer to the nearest-neighbor operation of (3) as the direct visual classifier, in the remainder of this work.

2.3. Indirect Contextual Classification System

The ICC system operates at the semantic level, representing images by vectors of concept counts $\mathcal{I} = (c_1, \ldots, c_L)^T$, where c_i is the number of feature vectors drawn from the i^{th} semantic concept. The count vector for the y^{th} image is drawn from a multinomial variable $\mathbf{T}$ of parameters $\boldsymbol{\pi}_y = (\pi_y^1, \ldots, \pi_y^L)^T$

$$P_{\mathbf{T}|Y}(\mathcal{I}|y; \boldsymbol{\pi}_y) \quad = \quad \frac{n!}{\prod_{k=1}^L c_k!} \prod_{j=1}^L (\pi_y^j)^{c_j}, \tag{4}$$

where π_y^i is the probability that an image feature vector is drawn from the i^{th} concept. We refer to the probability vector $\boldsymbol{\pi}_y$ that characterizes the y^{th} image as the *semantic multinomial* (SMN), and the space of all SMN's as the *semantic space*, denoted by $\mathcal{S}_L$. In the example of Fig. 1 this is a 371-dimensional vector space.

The indirect contextual classifier, then performs a nearest neighbor operation on the space $\mathcal{S}_L$, according to a similarity mapping $f : \mathcal{S}_L \to \{1, \ldots, D\}$ such that

$$f(\boldsymbol{\pi}) = \arg\max_y s(\boldsymbol{\pi}, \boldsymbol{\pi}_y) \tag{5}$$

where $\boldsymbol{\pi}$ is the query SMN, $\boldsymbol{\pi}_y$ the SMN of the y^{th} database image, and $s(\cdot, \cdot)$ an appropriate similarity function. We next describe, in more detail, a method for estimating SMNs, and a similarity function between them, which are compatible with the MPE decision rule.

2.3.1. *Semantic labeling system*

All SMNs $\boldsymbol{\pi}_i$ are learned with a semantic labeling system, which is implemented by computing posterior concept probabilities given the observed feature vectors

$$\pi_w = P_{W|\mathbf{X}}(w|\mathcal{I}). \tag{6}$$

A semantic class density $P_{\mathbf{X}|W}(\mathbf{x}|w)$ is learned for each concept w from the set $\mathcal{D}_w$ of all training images labeled with the w^{th} label in $\mathcal{L}$. This is based on a *hierarchical estimation* procedure [10], which estimates semantic class densities directly from the image densities used, for DVC, in (1). In this way, it is guaranteed that both the visual representation and the visual classification architecture used by the DVC and ICC systems are identical.

Given an image $\mathcal{I} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ the posterior concept probabilities of (6) are computed by combining (2) and Bayes rule, assuming a uniform prior concept distribution $P_W(w)$.

2.3.2. *Similarity function*

The similarity between SMNs $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$ is measured by the Kullback-Leibler divergence

$$s_{KL}(\boldsymbol{\pi}, \boldsymbol{\pi}') = KL(\boldsymbol{\pi}||\boldsymbol{\pi}') = \sum_{i=1}^L \pi_i \log \frac{\pi_i}{\pi'_i}. \tag{7}$$

This can be seen as the asymptotic limit of (3), when Y is uniformly distributed, guaranteeing consistency with the similarity function used for DVC.

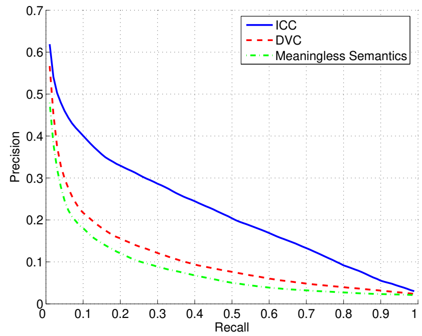


Fig. 2. Precision-recall curves achieved with ICC and DVC on *Corel50*. Also shown is the precision-recall with a meaningless semantic space.

Table 1. Contextual gains of ICC over DVC on all datasets.

Database	**Chance** (MAP)	**DVC** (MAP)	**ICC** (MAP)	**% CG**
Corel50	0.0200	0.1067	0.2259	111.73
Corel15	0.0667	0.2176	0.2980	36.95
Flickr18	0.0556	0.1373	0.2134	55.47

3. EXPERIMENTAL EVALUATION

In all experiments, the semantic space was learned from the Corel database used in [8, 11]. This dataset, henceforth referred to as *Corel50*, consists of $5,000$ images from 50 Corel Stock Photo CDs, of which 4500 images were used to learn the semantic space. Each image is labeled with 1-5 semantic concepts, from a set of 371 concepts, leading to a 371-dimensional semantic simplex. All images were converted from RGB to the YBR color space. Image observations were derived from 8×8 patches obtained with a sliding window, moved in a raster-scan fashion. A feature transformation was applied to this space by computing the 8×8 discrete cosine transform (DCT) of the three color components of each patch.

To evaluate classification performance of DVC and ICC systems, we carried out tests on three databases. First, the 4500 training images from *Corel50* served as the *retrieval database* and the remaining 500 as the database of *query images*. Next, we considered two databases *Corel15, Flickr18* where both the query and retrieval set contained concepts unknown to the semantic labeling system, that is they were composed of concepts from *outside the trained semantic space*. *Corel15*, comprised of $1,500$ images from 15 previously unused Corel CDs and *Flickr18* was collected from `www.flickr.com`, containing 1800 images divided into 18 classes according to the manual annotations provided by the online users. In both cases, 20% of randomly selected images served as *queries* and the remaining 80% as the *retrieval database*.

The performance was measured with precision-recall (PR) curves and mean average precision (MAP) [11]. The *contextual gain* of the ICC system was measured by

$$CG = \frac{MAP_{ICC} - MAP_{DVC}}{MAP_{DVC}} \times 100\%. \quad (8)$$

3.1. Contextual Gain

Table 1 summarizes the contextual gains of ICC over DVC, for all datasets considered. It is clear that ICC significantly outperforms DVC, the average contextual gain being of 111.73% for *Corel50*. Even outside the semantic space, the performance of the ICC system supersedes that of DVC system. In the case of *Flickr18* the gain is of 55.47%, and for *Corel15* of 36.95%. Since the visual representation and classification architecture are identical for the two approaches, this is strong indication that *there is a contextual gain*.

Fig. 2 also presents the PR curves obtained on *Corel50* with DVC and ICC. It can be seen that the precision of ICC is significantly higher than that of DVC, at all levels of recall. The benefits of contextual classification are also illustrated by Fig. 3, where we present some query results, under both DVC and ICC. Note that, for the example of Figure 1, the arch like structure of the locomotive rooftop is indeed a dominant feature for visual similarity: three of the five matches are images of bridges (using DVC). Nevertheless, the contextual correlations visible in the SMN of Figure 1, allow the ICC system to favor the correct locomotive interpretation.

To further investigate the contextual gain hypothesis we performed an experiment, using ICC with a semantically meaningless space. This was achieved by replicating the ICC experiments with random image groupings. That is, instead of a semantic space composed of concepts like “sky” (learned from images containing sky), we created a “semantic space” of nameless concepts learned from random collections of images. Fig. 2 compares (on *Corel50*) the PR obtained with ICC on this “meaningless semantic space”, with the previous results of DVC and ICC. It is clear that, *in the absence of semantic structure, ICC has very poor performance, and is clearly inferior to DVC*. This is further evidence that the source of the contextual gain of Table 1 are the contextual correlations of the underlying (meaningful) semantic space.

3.2. Growth rate of the contextual gain

Having established the existence of a contextual gain, we next study its dependence on two factors: the accuracy of the underlying DVC and the number of informative semantic dimensions.

3.2.1. Direct visual space

Since the visual representation is based on a subspace of the 192-dimensional space of DCT coefficients, it is possible to control the accuracy of DVC by simply varying the dimension of this subspace. As the number of features decreases, the performance of DVC tends to degrade. Fig. 4 (a) shows the MAP score for both DVC and ICC as the subspace dimension varies from 4 to 64. Note that the contextual gain is positive and increases with the accuracy of the visual classifiers. This is more clear in Fig. 4 (b) which shows the MAP of ICC as a function of that of DVC. The dashed blue line traces the set of points where the two MAPs are identical, i.e. where $CG = 0$. Notice that the slope of the curve is greater than 1, which implies the contextual gain has a positive growth rate with the accuracy of DVC. This suggests that although the accuracy of DVC is an important parameter for image classification, it should be possible to design good classification systems with less than perfect visual classifiers.

3.2.2. Informative dimensions of the semantic space

With respect to the impact of the structure the semantic space on the contextual gain, retrieval was performed for different numbers of dimensions of the semantic space. Semantic spaces of k dimensions, were produced by ordering the semantic feature by the variance of their posterior probabilities, and selecting the k of largest variance, (for k ranging from 0 to 371). Fig. 4 (c) shows a plot of the normalized MAP score (normalized by the maximum MAP for a given database) as a function of semantic space dimensions. The

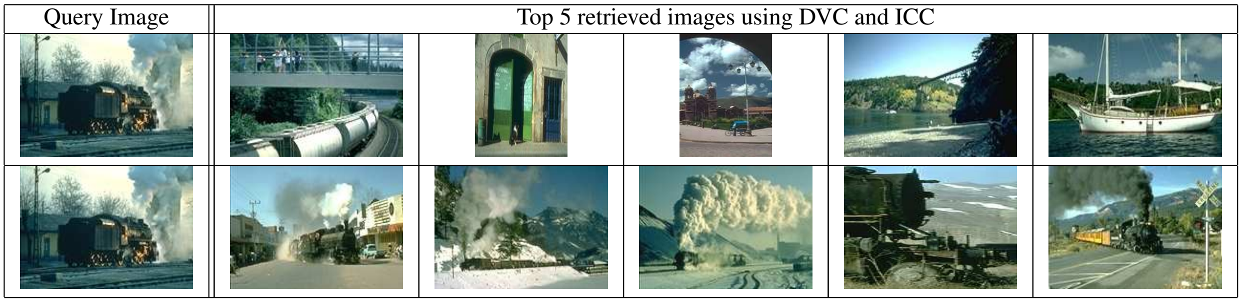
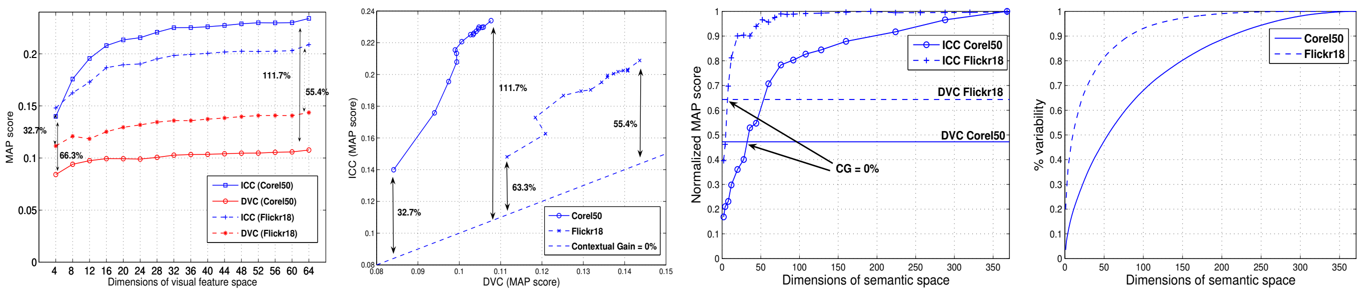


Fig. 3. Some examples where ICC performs better than DVC. The second row shows the images retrieved by ICC.



(a) (b) (c) (d)

Fig. 4. (a) MAP of DVC and ICC for *Corel50* and *Flickr18* as the accuracy of the underlying visual space varies. Also shown are the contextual gain at two extremes. (b) Contextual gain as a function of the accuracy of DVC. (c) Normalized MAP scores for ICC as a function of the number of semantic features. (d) % variability as explained by semantic concepts sorted according to the variance of their posterior probabilities.

plot also shows the normalized DVC score for both datasets. Notice that the contextual gain is positive for a semantic space with as little as 12(32) dimensions for *Flickr18*(*Corel50*). However, there is a saturation effect, i.e. not all 371 semantic concepts are equally informative. This is explained by Fig. 4 (d) which shows the variance of the posterior probabilities of the 371 semantic concepts. In particular, the MAP score saturated faster on *Flickr18* than on *Corel50* as almost all the variability is explained by around 100 concepts for *Flickr18* and more than 200 for *Corel50*. This shows that contextual correlations only help if the concepts are informative to start with.

4. DISCUSSION

In this work, we presented the first *systematic* study of the contextual gain hypothesis, i.e. that the ICC strategy outperforms the classical strategy of DVC. This study was based on a relatively simple classification architecture, which we do not claim to be the ultimate solution for image classification, but exhibits two properties of interest: 1) a unified architecture for both DVC and the visual component of ICC, which makes all performance gains attributable to the classification strategy, and 2) simplicity of implementation, which allowed us to control parameters, such as the accuracy of DVC, in a systematic and fine-grained manner. It produced a number of observations that, we believe, are of importance. The first was strong evidence in *support of the existence of a contextual gain*. This gain was consistent across various databases, and happened even when the images to classify depicted concepts not known to the semantic labeling system (outside the semantic space). Second, *the contextual gain appears to have a positive growth rate* with the accuracy of underlying DVC. Third, *contextual gains appear to be very easy to obtain*, as a positive contextual gain required, at most, 12 semantic features for *Flickr18*. Fourth, *contextual gain increases with the number of informative dimensions of the semantic space*. All these observations suggest that, while the improvement of DVC is an important direction of research for the advancement of image classification, it should be possible to design highly accurate recognizers with less than perfect visual classifiers.

5. REFERENCES

[1] I. Biederman, "On the semantics of a glance at a scene," *Perceptual organization*, pp. 213–263, 1981.

[2] D. Cox, E. Meyers, and P. Sinha, "Contextually evoked object-specific responses in human visual cortex," *Science*, vol. 304, no. 5667, pp. 115–117, 2004.

[3] L. Wolf and S. Bileschi, "A critical view of context," *International Journal of Computer Vision*, vol. 69, no. 2, pp. 251–261, 2006.

[4] A. Torralba, K.P. Murphy, and W.T. Freeman, "Contextual models for object detection using boosted random fields," *Advances in Neural Information Processing Systems*, 2004.

[5] P. Carbonetto, N. de Freitas, and K. Barnard, "A statistical model for general contextual object recognition," in *ECCV*, 2004.

[6] S. Kumar and M. Hebert, "Discriminative random fields: a discriminative framework for contextual interaction in classification," *ICCV*, pp. 1150–1157, 2003.

[7] N. Vasconcelos, "Minimum probability of error image retrieval," vol. 52, no. 8, August 2004.

[8] Nikhil Rasiwasia, Nuno Vasconcelos, and Pedro J. Moreno, "Query by semantic example.," in *CIVR*, 2006, pp. 51–60.

[9] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley and Sons, 2001.

[10] N. Vasconcelos, "Image indexing with mixture hierarchies," in *IEEE Computer Vision and Pattern Recognition Conf.*, Hawaii, 2001.

[11] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *CVPR*, Washington DC, 2004.