How Realistic is Photorealistic?

Siwei Lyu and Hany Farid
Department of Computer Science
Dartmouth College
Hanover, NH 03755
Email: {lyu,farid}@cs.dartmouth.edu

***Abstract*— Computer graphics rendering software is capable of generating highly photorealistic images that can be impossible to differentiate from photographic images. As a result, the unique stature of photographs as a definitive recording of events is being diminished (the ease with which digital images can be manipulated is, of course, also contributing to this demise). To this end, we describe a method for differentiating between photorealistic and photographic images. Specifically, we show that a statistical model based on first- and higher-order wavelet statistics reveals subtle but significant differences between photorealistic and photographic images.**

I. Introduction

Sophisticated computer graphics rendering software can generate remarkably photorealistic images. Though it may take some effort, photorealistic images can be created that are nearly impossible to differentiate from photographic images. And as the rendering technology improves, photorealistic images will become increasingly easier to generate and more realistic.

This technology is already having direct implications on our society. For example, in 1996 the United States Congress passed The Child Pornography Prevention Act which, in part, prohibited any image that *appears to be* or *conveys the impression of* someone under 18 engaged in sexually explicit conduct. This law made illegal computer generated pictures that only appear to show minors involved in sexual activity. In 2002, however, the United States Supreme Court struck down this law in their 6-3 ruling in *Ashcroft v. Free Speech Coalition* - the court said language in the 1996 child pornography law was unconstitutionally vague and far-reaching. This ruling makes it considerably more difficult for law enforcement agencies to prosecute child pornography crimes, since it is always possible to claim that any image is computer generated.

If we are to have any hope that photographs will again hold the unique stature of being a definitive recording of events, we must develop technology that can differentiate between photographic and photorealistic images. There has been some work in evaluating the photorealism of computer graphics rendered images from a human perception point of view (e.g., [10], [9], [11]). To our knowledge, however, no computational techniques exist to differentiate between photographic and photorealistic images (a method for differentiating between photographic and (non-realistic) graphical icons was proposed in [1]). Related work, though probably not directly applicable, include techniques to differentiate between city and landscape images [16], [14], in-door and out-door images [13], and photographs and paintings [4].

In this paper we describe a statistical model for photographic images that is built upon a wavelet-like decomposition. The model consists of first- and higher-order statistics that capture regularities that are inherent to photographic images. We then show that this model can be used to differentiate between photographic and photorealistic images - from a database of $40,000$ photographic and $6,000$ photorealistic images, we correctly classify approximately 67% of the photographic images while only mis-classifying approximately 1% of the photorealistic images. We have previously used a similar technique to detect messages hidden within digital images (steganography) [7], [8].

II. Statistical Model

The decomposition of images using basis functions that are localized in spatial position, orientation, and scale (e.g., wavelet) have proven extremely useful in image compression, image coding, noise removal, and texture synthesis. One reason is that such decompositions exhibit statistical regularities that can be exploited. The image decomposition employed here is based on separable quadrature mirror filters (QMFs) [15], [18], [12]. As illustrated in Figure 1, this decomposition splits the frequency space into multiple scales, and orientations (a vertical, a horizontal, and a diagonal subband). For a color (RGB) image, the decomposition is applied independently to each color channel. The resulting vertical, horizontal, and diagonal subbands for scale i are denoted

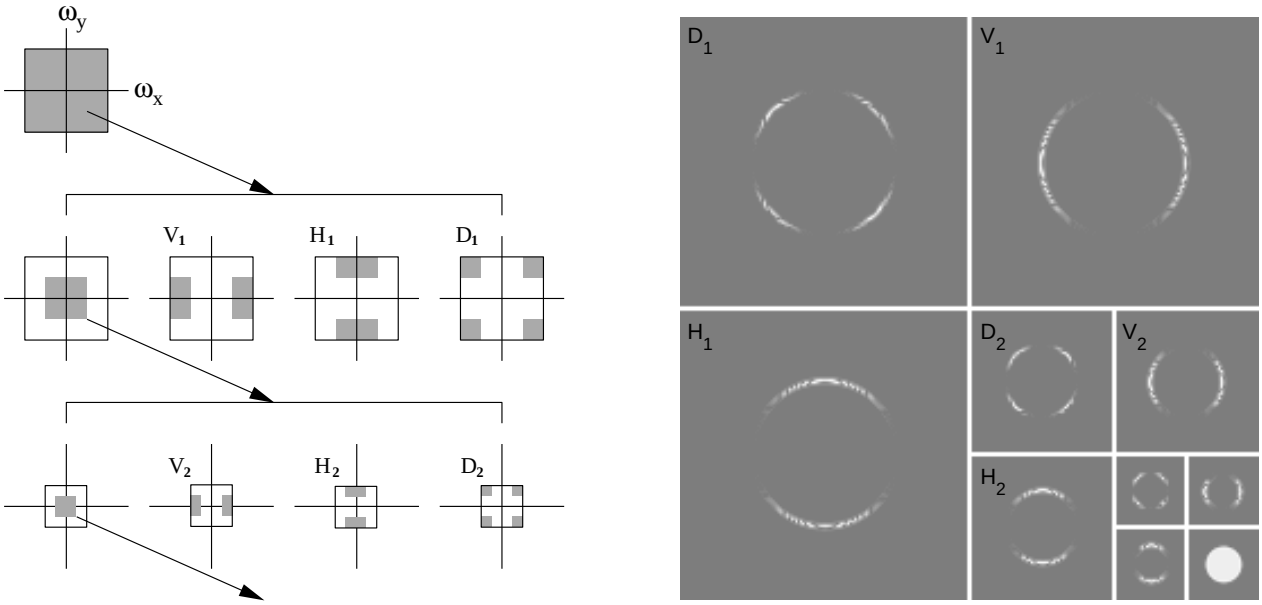

Fig. 1: Shown on the left is an idealized multi-scale and orientation decomposition of frequency space. Shown, from top to bottom, are levels 0, 1, and 2, and from left to right, are the low-pass, vertical, horizontal, and diagonal subbands. Shown on the right is the magnitude of a multi-scale and orientation decomposition of a "disc" image.

as $V_i^c(x,y)$, $H_i^c(x,y)$, and $D_i^c(x,y)$ respectively, where $c \in \{r,g,b\}$.

Wavelet subband coefficients for natural images typically follow a distribution which is well modeled by a generalized Laplacian $P(x) = \frac{1}{Z}e^{-|x/s|^p}$, where s,p are the density parameters, and Z is a normalizing constant [2]. This family of densities are characterized by a sharp peak at zero and large symmetric tails. An intuitive explanation for this is that natural images typically contain large smooth regions and abrupt transitions (e.g., edges). The smooth regions, though dominant, produce small coefficients near zero, while the transitions generate large coefficients. In our statistical model, instead of directly estimating the generalized Lapalacian distribution, a simpler approach is taken to characterize these marginal distributions. More specifically, the first four order statistics (mean, variance, skewness, and kurtosis) of the subband coefficient histograms at each orientation, scale, and color channel are collected. These statistics form the first half of our statistical model.

While these statistics describe the basic coefficient distributions, they are unlikely to capture the strong correlations that exist across space, orientation, and scale [2], [6]. For example, salient image features such as edges tend to orient spatially in certain direction and extend across multiple scales. These image features result in substantial local energy across many scales, orientations, and spatial locations. The local energy can be roughly measured by the magnitude of the decomposition coefficient. As such, a strong coefficient in a horizontal subband may indicate that its left and right spatial neighbors in the same subband will also have a large value. Similarly, if there is a coefficient with large magnitude at scale i, it is also very likely that its "parent" at scale $i+1$ will also have a large magnitude.

In order to capture some of these higher-order statistical correlations, we collect a second set of statistics that are based on the errors in a linear predictor of coefficient magnitude [2]. For the purpose of illustration, consider first a vertical band of the green channel at scale i, $V_i^g(x,y)$. A linear predictor for the magnitude of these coefficients in a subset [1] of all possible spatial, orientation, scale, and color neighbors is given by:

$$\begin{aligned} |V_i^g(x,y)| &= w_1|V_i^g(x-1,y)| + w_2|V_i^g(x+1,y)| \\ &+ w_3|V_i^g(x,y-1)| + w_4|V_i^g(x,y+1)| \\ &+ w_5|V_{i+1}^g(x/2,y/2)| + w_6|D_i^g(x,y)| \\ &+ w_7|D_{i+1}^g(x/2,y/2)| + w_8|V_i^r(x,y)| \\ &+ w_9|V_i^b(x,y)|, \end{aligned} \quad (1)$$

where $|\cdot|$ denotes absolute value and w_k are the scalar weights. This linear relationship can be expressed more compactly in matrix form as:

$$\vec{v} = Q\vec{w}, \quad (2)$$

[1]The particular choice of neighbors was motivated by the observations of [2] and modified to include non-casual neighbors.

where $\vec{v}$ contains the coefficient magnitudes of $V_i^g(x,y)$ strung out into a column vector (to reduce sensitivity to noise, only magnitudes greater than 1 are considered), the columns of the matrix Q contain the neighboring coefficient magnitudes as specified in Equation (1), and $\vec{w} = (w_1 \ \ ... \ \ w_9)^T$. The weights $\vec{w}$ are determined by minimizing the following quadratic error function:

$$E(\vec{w}) \ = \ [\vec{v} - Q\vec{w}]^2. \quad (3)$$

This error function is minimized by differentiating with respect to $\vec{w}$:

$$\frac{dE(\vec{w})}{d\vec{w}} \ = \ 2Q^T(\vec{v} - Q\vec{w}), \quad (4)$$

setting the result equal to zero, and solving for $\vec{w}$ to yield:

$$\vec{w} \ = \ (Q^TQ)^{-1}Q^T\vec{v}. \quad (5)$$

Given the large number of constraints (one per pixel) in only nine unknowns, it is generally safe to assume that the 9×9 matrix Q^TQ will be invertible.

Given the linear predictor, the log error between the actual coefficient and the predicted coefficient magnitudes is:

$$\vec{p} \ = \ \log(\vec{v}) - \log(|Q\vec{w}|), \quad (6)$$

where the $\log(\cdot)$ is computed point-wise on each vector component. As with the coefficient statisics, mean, variance, skewness, and kurtosis of this error distribution are collected. This process is repeated for scales $i = 1, ..., n-1$, and for the subbands V_i^r and V_i^b, where the linear predictors for these subbands are of the form:

$$\begin{aligned} |V_i^r(x,y)| &= w_1|V_i^r(x-1,y)| + w_2|V_i^r(x+1,y)| \\ &+ w_3|V_i^r(x,y-1)| + w_4|V_i^r(x,y+1)| \\ &+ w_5|V_{i+1}^r(x/2,y/2)| + w_6|D_i^r(x,y)| \\ &+ w_7|D_{i+1}^r(x/2,y/2)| + w_8|V_i^g(x,y)| \\ &+ w_9|V_i^b(x,y)|, \end{aligned} \quad (7)$$

and

$$\begin{aligned} |V_i^b(x,y)| &= w_1|V_i^b(x-1,y)| + w_2|V_i^b(x+1,y)| \\ &+ w_3|V_i^b(x,y-1)| + w_4|V_i^b(x,y+1)| \\ &+ w_5|V_{i+1}^b(x/2,y/2)| + w_6|D_i^b(x,y)| \\ &+ w_7|D_{i+1}^b(x/2,y/2)| + w_8|V_i^r(x,y)| \\ &+ w_9|V_i^g(x,y)|. \end{aligned} \quad (8)$$

A similar process is repeated for the horizontal and diagonal subbands. As an example, the predictor for the green channel takes the form:

$$\begin{aligned} |H_i^g(x,y)| &= w_1|H_i^g(x-1,y)| + w_2|H_i^g(x+1,y)| \\ &+ w_3|H_i^g(x,y-1)| + w_4|H_i^g(x,y+1)| \\ &+ w_5|H_{i+1}^g(x/2,y/2)| + w_6|D_i^g(x,y)| \\ &+ w_7|D_{i+1}^g(x/2,y/2)| + w_8|H_i^r(x,y)| \\ &+ w_9|H_i^b(x,y)|, \end{aligned} \quad (9)$$

and

$$\begin{aligned} |D_i^g(x,y)| &= w_1|D_i^g(x-1,y)| + w_2|D_i^g(x+1,y)| \\ &+ w_3|D_i^g(x,y-1)| + w_4|D_i^g(x,y+1)| \\ &+ w_5|D_{i+1}^g(x/2,y/2)| + w_6|H_i^g(x,y)| \\ &+ w_7|V_i^g(x,y)| + w_8|D_i^r(x,y)| \\ &+ w_9|D_i^b(x,y)|. \end{aligned} \quad (10)$$

For the horizontal and diagonal subbands, the predictor for the red and blue channels are determined in a similar way as was done for the vertical subbands, Equations (7)-(8). For each oriented, scale and color subband, a similar error metric, Equation (6), and error statistics are computed.

For a multi-scale decomposition with scales $i = 1, ..., n$, the total number of basic coefficient statistics is $36(n-1)$ ($12(n-1)$ per color channel), and the total number of error statistics is also $36(n-1)$, yielding a grand total of $72(n-1)$ statistics. These statistics form the feature vector to be used to discriminate between photorealistic and photographic images.

III. Classification

From the measured statistics of a training set of images labeled as photorealistic or photographic, our goal is to build a classifier that can determine to which category a novel test image belongs.

To this end, linear discrimination analysis (LDA) is a widely used classification algorithm [5]. In a two-class LDA a one-dimensional linear subspace is found such that when the features are projected onto this subspace, the within-class scatter is minimized while the between-class scatter is maximized. LDA is attractive because of its general effectiveness and simplicity (the classifier is built using a closed-form generalized eigenvector solution). The drawback of LDA is that the classification surface is constrained to be linear.

Support vector machines (SVM) afford a more flexible non-linear classification surface [17]. Within this family of classifiers there are both linear and non-linear SVMs. A linear SVM is similar to an LDA, the difference being in the objective function that is minimized. A non-linear SVM extends a linear SVM by using a kernel function to

map the training exemplars into a higher (possibly infinite) dimensional space. While affording a more flexible classifier, the construction of a non-linear SVM is no longer closed-form, but requires an iterative numerical optimization.

We employed both LDA and a non-linear SVM for the purposes of distinguishing between photorealistic and photographic images.

IV. Results

Shown in Figures 2 and 3 are several images taken from a database of $40,000$ photographic and 6,000 photorealistic images[2]. All of the images consist of a broad range of indoor and outdoor scenes, and the photorealistic images were rendered using a number of different software packages (e.g., 3D Studio Max, Maya, Soft-Image 3D, PovRay, Lightwave 3D and Imagine). All of the images are color (RGB), JPEG compressed (with an average quality of 90%), and typically on the order of 600×400 pixels in size.

From this database of $46,000$ images, statistics as described in Section II were extracted. To accommodate different image sizes, only the central 256×256 region of each image was considered. For each image region, a four-level three-orientation QMF pyramid [3] was constructed for each color channel, from which a 216-dimensional feature vector (72 per color channel) of coefficient and error statistics was collected.

From the $46,000$ feature vectors, $32,000$ photographic and $4,800$ photorealistic feature vectors were used to train both an LDA and a non-linear SVM [4]. The remaining feature vectors were used to test the classifiers. In the results presented here, the training/testing split was done randomly. We report, in Table I, the classification accuracy averaged over 100 such splits. With an 0.8% false-negative rate (a photorealistic image classified as photographic), the LDA correctly classified 54.6% of the photographic images. A non-linear SVM had better performance, correctly classifying 66.8% of the photographic images, with a 1.2% false-negative rate (the variances over the 100 splits was 3.46% and 0.09%, respectively). Note that in both cases this testing accuracy was fairly close to the training accuracy, suggesting that the classifiers generalized.

	training		testing	
	LDA	SVM	LDA	SVM
photographic	58.7	70.9	54.6	66.8
photorealistic	99.4	99.1	99.2	98.8

TABLE I: Classification results using LDA and SVM. Shown are the average accuracies (in percent) over 100 random training/testing splits of the database of $40,000$ photographic and $6,000$ photorealistic images.

We next wondered which images were most easy and most difficult to classify. Specifically, images that are easy to classify are those that are far from the separating classification surface, and those that are hard to classify are near, or on the wrong side of, the classification surface. Shown in Figures 4 and 5 are eight photographic images and eight photorealistic images, respectively, that were easily classified under the non-linear SVM. Shown in Figure 6 are eight photographic images, furthest away from the classification surface, that were incorrectly classified. Shown in Figure 7 are eight incorrectly classified photorealistic images, furthest away from the classification surface.

We further tested the RBF SVM classifier on a novel set of fourteen images (7 photographic, 7 photorealistic) from the website `www.fakeorfoto.com`. Shown in Figure 9 are the fourteen images with the correctly classified photographic images in the top row, and the correctly classified photorealistic images in the middle row. Shown in the bottom row are three incorrectly classified photographic images (left) and two incorrectly classified photorealistic images (right).

We wondered which set of statistics, coefficient or error, were most crucial for the classifier. Shown in Figure 8 is the accuracy of the classifier plotted against the number and category of feature for the LDA classifier [5]. We began by choosing the single feature, out of the 216

[2]The photographic images were downloaded from `www.freefoto.com`, the photorealistic images were downloaded from `www.raph.com` and `www.irtc.org`.

[3]We employed a 9-tap QMF filter as the basis of the multi-scale multi-orientation image decomposition. The low-pass, l, and high-pass, h, filters are given by:

$$\begin{aligned} l &= [0.02807382 \quad -0.060944743 \quad -0.073386624 \\ &\quad 0.41472545 \quad 0.7973934 \quad 0.41472545 \\ &\quad -0.073386624 \quad -0.060944743 \quad 0.02807382] \\ h &= [0.02807382 \quad 0.060944743 \quad -0.073386624 \\ &\quad -0.41472545 \quad 0.7973934 \quad -0.41472545 \\ &\quad -0.073386624 \quad 0.060944743 \quad 0.02807382]. \end{aligned}$$

We also have experimented with both Laplacian and steerable pyramid decompositions. Results from a steerable pyramid (with eight orientation subbands) were similar to the results using a QMF pyramid (which use only three orientation subbands). The Laplacian pyramid generally gave poor results. So while it seems that oriented subbands are necessary, it also seems that a finer orientation tuning is not necessary for this particular task.

[4]We employed the SVM algorithm implemented in LIBSVM [3], along with an RBF kernel.

[5]This analysis was performed only on the LDA because the computational cost of retraining $23,220 = 216 + \cdots + 1$ non-linear SVMs is prohibitive. We expect the same pattern of results for the non-linear SVM.

possible coefficient and error features, that gives the best classification accuracy. This was done by building 216 classifiers each based on a single feature, and choosing the feature that yields the highest accuracy (the feature was the variance in the error of the green channel's diagonal band at the second scale). We then choose the next best feature from the remaining 215 components. This process was repeated until all features were selected. The solid line in Figure 8 is the accuracy as a function of the number of features. The white and gray regions correspond to error and coefficient features, respectively. That is, if the feature included on the i^{th} iteration is a coefficient then we denote that with a vertical gray line at the i^{th} position on the horizontal axis. Note that the coefficient and error statistics are interleaved, showing that both sets of statistics are important for classification.

And finally, we attempted to retrain the non-linear SVM with random class labels assigned to the training images. The rationale for this was to ensure that the statistical model and classifier are discriminating on fundamental differences between photographic and photorealistic images, and not on some artifact. To this end, we expect a random class assignment to lead to significantly worse classification accuracy. We generated ten different training sets containing $5,000$ randomly selected photographic images and $5,000$ photorealistic images. One-half of these images were randomly assigned to the photographic class and the other half were assigned to the photorealistic class. We then trained non-linear SVM classifiers on these training sets and tested them on the testing sets as used in our experiment described above. The best performance across the ten training sets was 27.6% correctly classified photographic images, with a 1.4% false-negative rate. Note that this is significantly worse than the 66.8% detection accuracy when the correct training labels were used. This result indicates that our statistical model and classifier are discriminating on fundamental statistical differences between photographic and photorealistic images.

V. Discussion

We have described a statistical model for photographic images consisting of first- and higher-order wavelet statistics. This model seems to capture regularities that are inherent to photographic images. We have also shown that this model, coupled with either an LDA or a non-linear SVM, can be used to differentiate between photorealistic and photographic images. It is interesting to see that even though photorealistic images can be perceptually indistinguishable from photographic images, their underlying statistics can still be significantly different. These techniques are also likely to have important applications in the growing field of digital forensics.

There are, of course, several possible extensions to this work. We expect that these techniques can be extended to differentiate between synthetically generated and natural voice signals and video streams. And, as in earlier work [8] we expect a one-class SVM, that only requires training from photographic images, to simplify the classifier training.

Finally we note that it is not immediately obvious that a photorealistic image could be altered to match the expected higher-order statistics of photographic images. The drawback of this, from a rendering point of view, is that these models don't necessarily give any insight into how one might render more photorealistic images. The benefit, from a digital forensic point of view, is that it is likely that this model will not be immediately vulnerable to counter-attacks. It is possible, of course, that counter-measures will be developed that can foil the classification scheme outlined here. The development of such techniques will in turn lead to better classification schemes, and so on.

Acknowledgments

This work was supported by an Alfred P. Sloan Fellowship, an NSF CAREER Award (IIS99-83806), an NSF Infrastructure Grant (EIA-98-02068), and under Award No. 2000-DT-CX-K001 from the Office for Domestic Preparedness, U.S. Department of Homeland Security (points of view in this document are those of the authors and do not necessarily represent the official position of the U.S. Department of Homeland Security).

References

[1] V. Athitsos, M.J. Swain, and C. Frankel. Distinguishing photographs and graphics on the world wide web. In *Workshop on Content-Based Access of Image and Video Libraries*, 1997.

[2] R.W. Buccigrossi and E.P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing*, 8(12):1688–1701, 1999.

[3] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001.

[4] F. Cutzu, R. Hammoud, and A. Leykin. Estimating the degree of photorealism of images: Distinguishing paintings from photographs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[5] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.

[6] J. Gluckman. On the use of marginal statistics of subband images. In *IEEE International Conference on Computer Vision*, Nice, France, 2003.

[7] S. Lyu and H. Farid. Detecting hidden messages using higher-order statistics and support vector machines. In *5th International Workshop on Information Hiding*, 2002.

[8] S. Lyu and H. Farid. Steganalysis using color wavelet statistics and one-class support vector machines. In *SPIE Symposium on Electronic Imaging*, 2004.

[9] A. McNamara. Evaluating realism. In *Perceptually Adaptive Graphics, ACM SIGGRAPH and Eurographics Campfire*, 2001.

[10] G.W. Meyer, H.E. Rushmeier, M.F. Cohen, D.P. Greenberg, and K.E. Torrance. An experimental evaluation of computer graphics imagery. *ACM Transactions on Graphics*, 5(1):30–50, 1986.

[11] P.M. Rademacher. *Measuring the Perceived Visual Realism of Images*. PhD thesis, UNC at Chapel Hill, 2002.

[12] E.P. Simoncelli and E.H. Adelson. *Subband image coding*, chapter Subband transforms, pages 143–192. Kluwer Academic, 1990.

[13] M. Szummer and R.W. Picard. Indoor-outdoor image classification. In *IEEE International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV*, pages 42–51, 1998.

[14] A. Torralba and A. Oliva. Semantic organization of scenes using discriminant structural templates. In *International Conference on Computer Vision*, 1999.

[15] P.P. Vaidyanathan. Quadrature mirror filter banks, M-band extensions and perfect reconstruction techniques. *IEEE ASSP Magazine*, pages 4–20, 1987.

[16] A. Vailaya, A.K. Jain, and H.-J. Zhang. On image classification: City vs. landscapes. *International Journal of Pattern Recognition*, (31):1921–1936, 1998.

[17] V. Vapnik. *The nature of statistical learning theory*. Spring Verlag, 1995.

[18] M. Vetterli. A theory of multirate filter banks. *IEEE Transactions on ASSP*, 35(3):356–372, 1987.

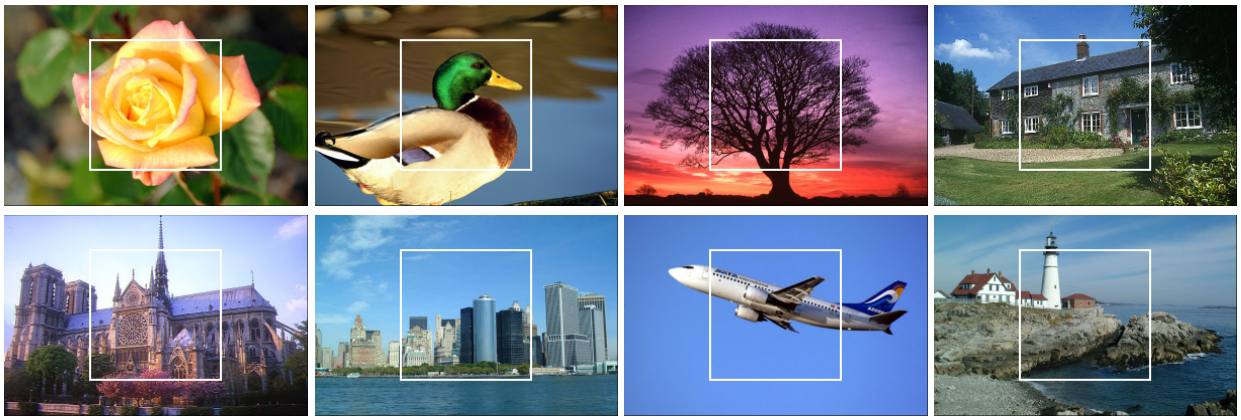

Fig. 2: Eight examples from a database of $40,000$ photographic images. The central 256×256 white boxes denote the region of the image from which statistics are measured.

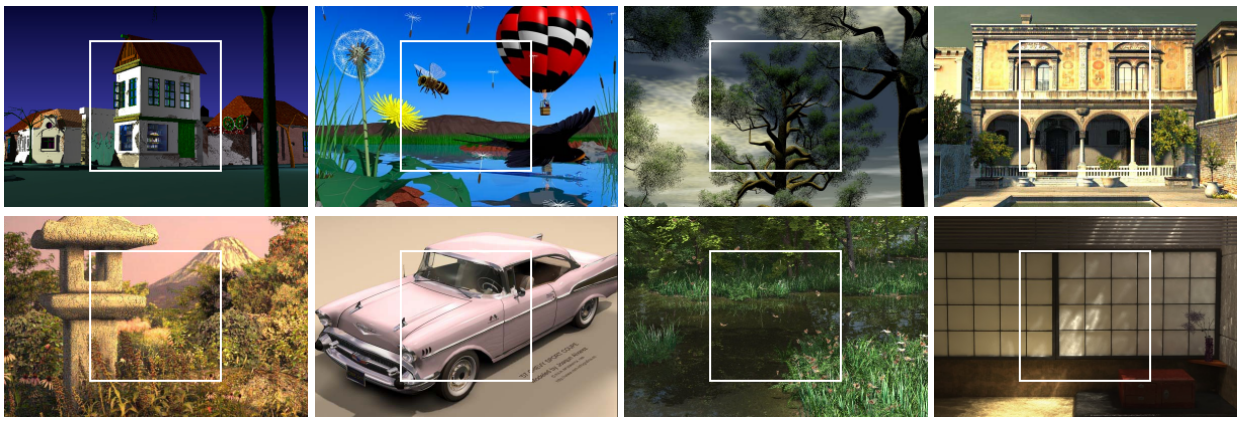

Fig. 3: Eight examples from a database of $6,000$ photorealistic images. The central 256×256 white boxes denote the region of the image from which statistics are measured.

Fig. 4: Easily classified photographic images.

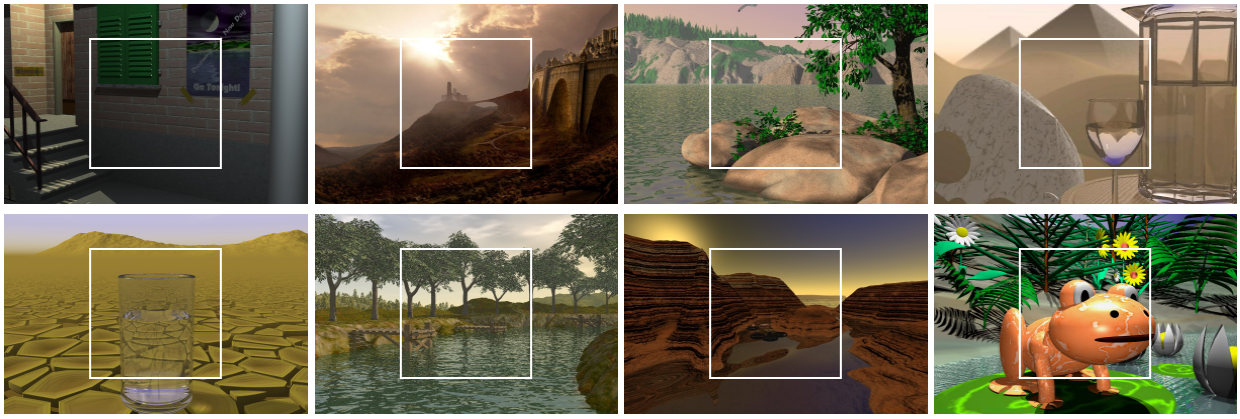

Fig. 5: Easily classified photorealistic images.

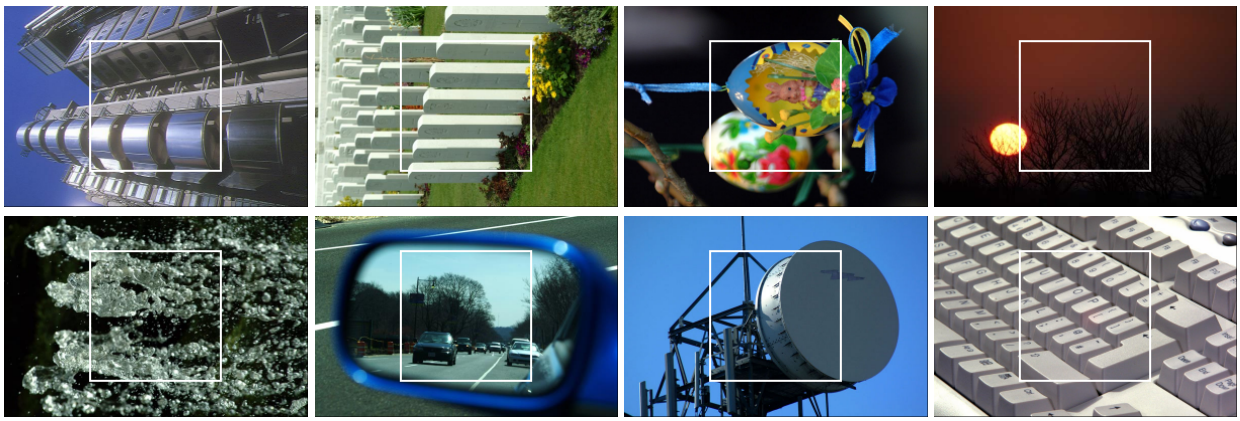

Fig. 6: Incorrectly classified photographic images.

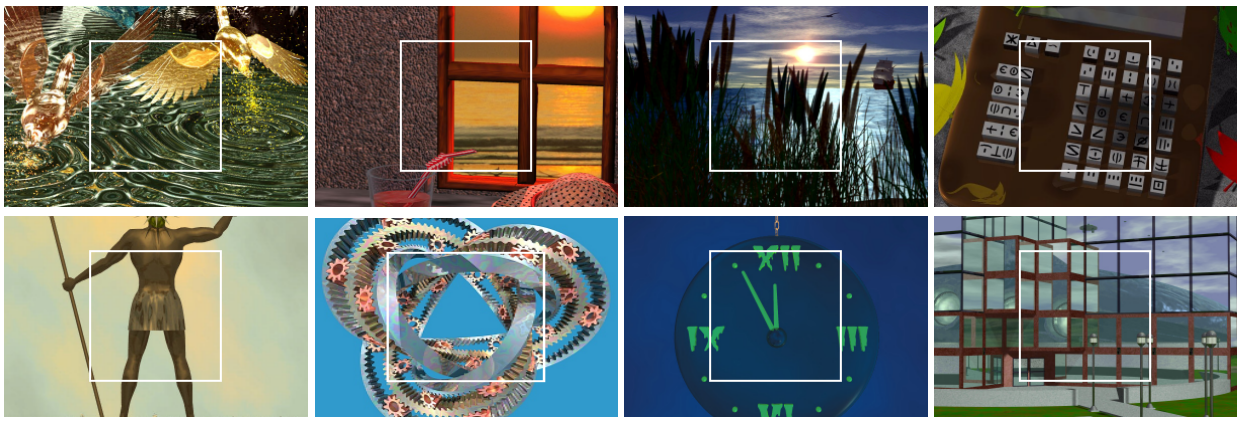

Fig. 7: Incorrectly classified photorealistic images.

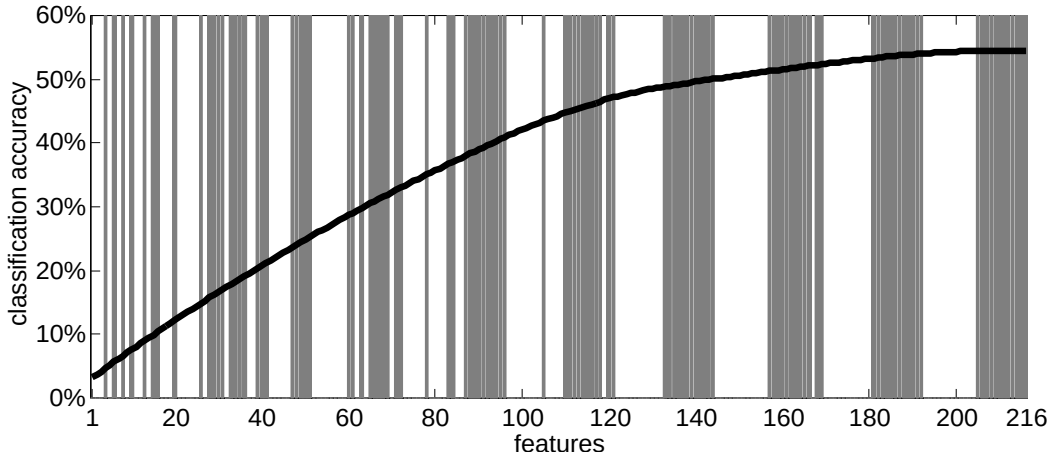

Fig. 8: Shown is the classification accuracy as a function of the number and category of feature for the LDA classifier. The white and gray regions correspond to error and coefficient features, respectively.

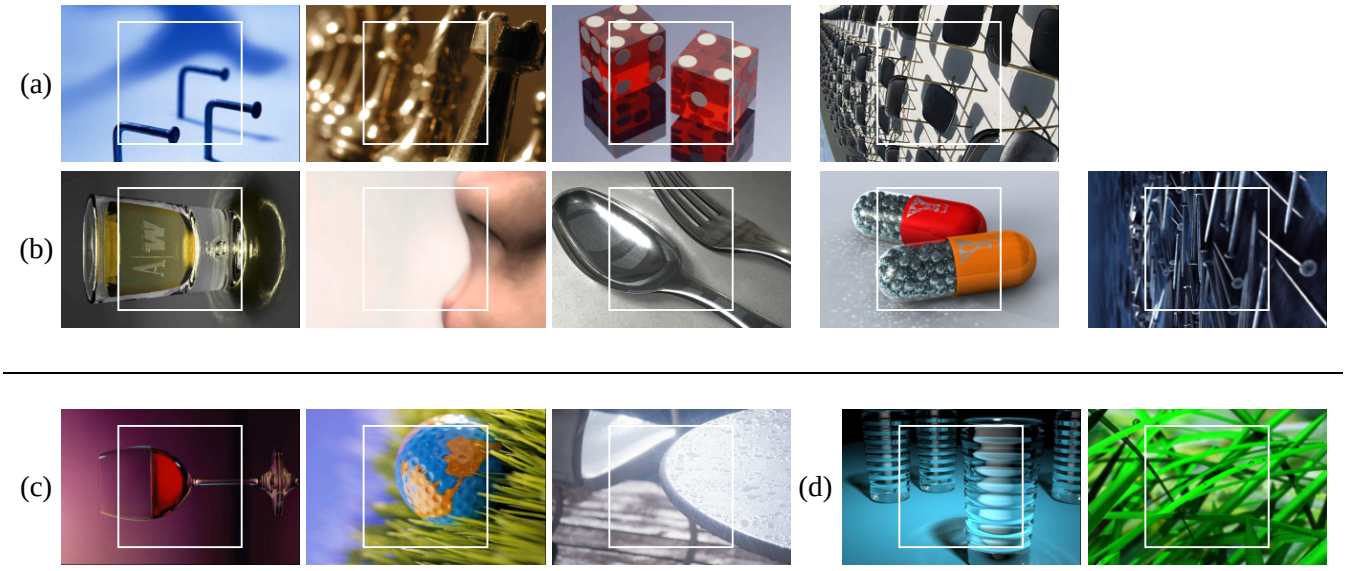

Fig. 9: Images from `www.fakeorfoto.com`. Shown in (a) and (c) are correctly and incorrectly classified photographic images, respectively. Shown in (b) and (d) are correctly and incorrectly classified photorealistic images, respectively.