Pergamon

Pattern Recognition, Vol. 30, No. 5, pp. 751–768, 1997
© 1997 Pattern Recognition Society. Published by Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0031-3203/97 $17.00+.00

PII: S0031-3203(96)00104-5

TEMPLATE MATCHING: MATCHED SPATIAL FILTERS AND BEYOND

R. BRUNELLI[†,*] and T. POGGIO[‡]

[†]Istituto per la Ricerca Scientifica e Tecnologica, I-38050 Povo, Trento, Italy
[‡]Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, U.S.A.

(*Received* 27 *November* 1995; *in revised form* 1 *July* 1996; *received for publication* 15 *July* 1996)

Abstract—Template matching by means of cross-correlation is common practice in pattern recognition in spite of its drawbacks. This paper reviews some results on how these shortcomings can be removed. Several techniques (Matched Spatial Filters, Synthetic Discriminant Functions, Principal Components Projections and Reconstruction Residuals) are reviewed and compared on a common task: locating eyes in a database of faces. New variants are also proposed and compared: least squares Discriminant Functions and the combined use of projections on eigenfunctions and the corresponding reconstruction residuals. Finally, approximation networks are introduced in an attempt to improve filter design by the introduction of nonlinearity. © 1997 Pattern Recognition Society. Published by Elsevier Science Ltd.

Template matching Correlation Neural networks Learning
HyperBF networks Principal components

1. INTRODUCTION

The detection and recognition of objects from their images, irrespective of their orientation, scale, and view, is a very important research subject in computer vision, if not computer vision itself. Several techniques have been proposed in the past to solve this challenging problem. In this paper we will focus on a subset of these techniques, those employing the idea of projection to match image patterns. The notion of Matched Spatial Filter (MSF) is a venerable one with a long history.[1] While by itself it cannot account for invariant recognition, it can be coupled to invariant mappings or signal expansions, and is therefore able to provide invariance to rotation and scaling in the image plane. In order to cope with more general variations of the object's views more sophisticated approaches have to be employed. Among them, the use of Synthetic Discriminant Functions[2–14] is one of the more promising so far developed. In these paper we will follow a path from MSF, to expansion matching through different variant of SDFs. Section 2 describes the basic properties of MSF, their optimality and their relation to the probability of misclassification. The generalization of MSF to a linear combination of example images is introduced next. Several shortcomings of the basic approach are outlined and a set of possible solutions is presented in the subsequent section. We discuss a relation of the resulting class of filters to nonorthogonal image expansion. A generalization to projections on multiple directions and the use of the projection residual for pattern matching is then investigated.[15–20] Finally, a more powerful, nonlinear framework is introduced in which template matching can be looked at as a problem of function approximation. Network architectures and training strategies are proposed within this new general framework.

* Author to whom correspondence should be addressed.

2. MATCHED SPATIAL FILTER

Template matching is extensively used in low-level vision tasks to localize and identify patterns in images. Two methods are commonly used:

1. Image subtraction: images are considered as vectors and the norm of their difference is considered as a measure of dissimilarity;
2. Correlation: the dot product of two images is considered as a measure of their similarity (it represents the angle between the images when they are suitably normalized and considered as vectors).

When the images are normalized to have zero average and unit norm, the two approaches give the same result. The usual implementation of the above methods relies on the Euclidean distance. Other distances can be used and some of them have better properties such as increased robustness to noise and minor deformations.[21] The next sections are mainly concerned with the correlation approach. The idea of image subtraction is introduced again in the more general nonlinear framework.

2.1. *Optimality*

One of the reasons for which template matching by correlation is commonly used is that correlation can be shown to be the optimal (according to a particular criterion) linear operation by which a deterministic signal corrupted by additive white Gaussian noise can be

Magna 2046

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

detected.[1] Let the signal be

$$g(x) = \phi(x) + \lambda(x), \tag{1}$$

where $\phi(x)$ is the original, uncorrupted, signal and $\lambda(x)$ is noise with power spectrum $S(\omega)$. The noise is assumed to be wide-sense stationary with zero average so that

$$E\{\lambda(x)\} = 0, \qquad E\{\lambda(x+\alpha)\lambda(x)\} = R(\alpha).$$

We assume that $\phi(x)$ is known and we want to establish its presence and location. To do so we apply to the process $g(x)$ a linear filter with impulse response $h(x)$ and system function $H(\omega)$. The resulting output is

$$z(x) = g(x) * h(x) = \int_{-\infty}^{\infty} g(x-\alpha)h(\alpha)\,\mathrm{d}\alpha \tag{2}$$

$$= z_\phi(x) + z_\lambda(x). \tag{3}$$

Using the convolution theorem for the Fourier transform we have that

$$z_\phi(x) = \int_{-\infty}^{\infty} \phi(x-\alpha)h(\alpha)\,\mathrm{d}\alpha$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi(\omega)H(\omega)\exp[\mathrm{i}\omega x]\,\mathrm{d}\omega.$$

We want to find $H(\omega)$ so as to maximize the following signal to noise ratio (SNR):

$$r^2 = \frac{|z_\phi(x_0)|^2}{E\{z_\lambda^2(x_0)\}}, \tag{4}$$

where x_0 is the location of the signal. The SNR represents the ratio of the filter responses at the uncorrupted signal and at the noise. It is defined at the true location of the signal (usually the correlation peak) therefore not taking into account the off-peak response of the filter.

Two cases of particular interest are those of white and colored noise:

White Noise: This type of noise is defined by the following condition:

$$S(\omega) = S_0,$$

which corresponds to a flat energy spectrum. The Schwartz inequality states that

$$\left|\int_a^b f(x)g(x)\,\mathrm{d}x\right|^2 \le \int_a^b |f(x)|^2 \int_a^b |g(x)|^2$$

and the equality holds iff $f(x) = k\tilde{g}(x)$ (we use $\tilde{\ }$ to represent complex conjugation). This implies the following bound for the signal to noise ratio r:

$$r^2 \le \frac{\int |\Phi(\omega)\exp[\mathrm{i}\omega x_0]|^2\,\mathrm{d}\omega \int |H(\omega)|^2\,\mathrm{d}\omega}{2\pi S_0 \int |H(\omega)|^2\,\mathrm{d}\omega}$$

and then

$$r^2 \le \frac{E_\Phi}{S_0},$$

where

$$E_\Phi = \frac{1}{2\pi}\int |\Phi(\omega)|^2\,\mathrm{d}\omega$$

represents the energy of the signal. From the Schwartz inequality the equality holds only if

$$H(\omega) = k\tilde{\Phi}(\omega)\exp[-\mathrm{i}\omega x_0].$$

The spatial domain version of the filter is simply the mirror image of the signal:

$$h(x) = k\phi(x_0 - x)$$

which implies that the convolution of the signal with the filter can be expressed as the cross-correlation with the signal (hence the name Matched Spatial Filter).

Colored Noise: If the noise has a nonflat spectrum $S(\omega)$ it is said to be colored. In this case the following holds:

$$2\pi z_\phi(x_0) = \int \Phi(\omega)H(\omega)\exp[\mathrm{i}\omega x]\,\mathrm{d}\omega,$$

$$|2\pi z_\phi(x_0)|^2 = \left|\int \frac{\Phi(\omega)}{\sqrt{S(\omega)}}\sqrt{S(\omega)}H(\omega)\exp[\mathrm{i}\omega x]\,\mathrm{d}\omega\right|^2$$

$$\le \int \frac{|\Phi(\omega)\exp[\mathrm{i}\omega x]|^2}{S(\omega)}\,\mathrm{d}\omega \times \int S(\omega)|H(\omega)|^2\,\mathrm{d}\omega,$$

hence

$$r^2 \le \frac{1}{2\pi}\int \frac{|\Phi(\omega)\exp[\mathrm{i}\omega x]|^2}{S(\omega)}\,\mathrm{d}\omega$$

with equality holding only when

$$\sqrt{S(\omega)}H(\omega) = k\frac{\tilde{\Phi}\exp[-\mathrm{i}\omega x_0]}{\sqrt{S(\omega)}}.$$

The main consequence of the color of noise is that the optimal filter corresponds to a modified version of the signal

$$H(\omega) = k\frac{\tilde{\Phi}\exp[-\mathrm{i}\omega x_0]}{S(\omega)},$$

which emphasizes the frequencies where the energy of the noise is smaller. The optimal filter can also be considered as a cascade of a whitening filter $S^{-1/2}(\omega)$ and the usual filter based on the transformed signal.

In the spatial domain, correlation amounts to projecting the signal $g(x)$ onto the available *template* $\phi(x)$. If the norm of the projected signal is not equal to that of the template, the value of the projection can be meaningless as the projection value can be large without implying that the two vectors are close in any reasonable sense. The solution is to compute the projection using normalized vectors. In particular, if versors are used, computing the projection amounts to computing the cosine of the angle formed by the two vectors, which is an effective measure of similarity. In vision tasks, vector normalization corresponds to adjusting the intensity scale so that the corresponding distribution has a given variance. Another

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.

useful normalization is to set the average value of the vector coordinates to zero. This operation corresponds to setting the average of the intensity distribution for images. These normalizations are particularly useful when modern cameras are used, as they usually operate with automatic gain level (acting on the scale of the intensity) and black level adjustment (acting as an offset on the intensity distribution).

2.2. Distorted templates

The previous analysis was focused on the detection of a deterministic signal corrupted by noise. An interesting extension is the detection of a signal belonging to a given distribution of signals.[2] As an example, consider the problem of locating the eyes in a face image. We do not know who's face it is so that we cannot select the corresponding signal (the eyes of that person). A whole set of different eyes could be available, possibly including the correct ones.

Let $\{\phi(x)\}$ denote the class of signals to be detected. We want to find the filter h which maximizes the SNR r^2 over the class of signals $\{\phi(x)\}$. The input signal $\phi(x)$ can be modeled as a sample realization of the stochastic process $\{\phi(x)\}$. The ensemble-average correlation function of the stochastic process is defined by

$$K_{\phi\phi}(x,y) = E_\phi\{\phi(x)\phi(y)\} \tag{5}$$

and represents the average over the ensemble of signals (and not over the coordinates of a signal). What we want to maximize is the ensemble average of the signal to noise ratio:

$$E_\phi\{r^2\} = \frac{E\{|z_\phi(x_0)|^2\}}{E\{z_\lambda^2(x_0)\}}. \tag{6}$$

Assume, without loss of generality, that x_0=0. The average SNR can then be rewritten as

$$E_\phi\{r^2\} = \frac{\int\int h(-x)h(-y)K_{\phi\phi}(x,y)\,dx\,dy}{\int\int h(-x)h(-y)K_{\lambda\lambda}(x,y)\,dx\,dy}, \tag{7}$$

where the ensemble autocorrelation function of the signal and noise have been used. The autocorrelation function of the white noise is proportional to a Dirac delta function:

$$K_{\lambda\lambda}(x,y) = N\delta(x-y) \tag{8}$$

so that the average signal to noise ratio can be rewritten as

$$E_\phi\{r^2\} = \frac{\int\int h(-x)h(-y)K_{\phi\phi}(x,y)\,dx\,dy}{N\int h(-x)^2\,dx}. \tag{9}$$

Pre-whitening operators can be applied as preprocessing functions when the assumption of white noise does not hold. The denominator of the RHS in equation (9) represents the energy of the filter and we can require it to be 1:

$$\int h(-x)^2\,dx = 1. \tag{10}$$

To optimize equation (9) we must maximize the numerator subject to the energy constraint of the filter. The ensemble autocorrelation function can be expressed in terms of the orthonormal eigenfunctions of the integral kernel $K_{\phi\phi}(x,y)$

$$K_{\phi\phi}(x,y) = \sum_i \lambda_i\psi_i(x)\psi_i(y), \tag{11}$$

where the λ_i are the corresponding eigenvalues. The filter function h can also be expanded in the same basis

$$h(-x) = \sum_i \omega_i\psi_i(x). \tag{12}$$

Using the inner product notation and the orthonormality of the $\psi_i(x)$ we can state the optimization problem as finding

$$\arg \max_{\sum_i \omega_i^2=1} \sum_i \lambda_i(h\cdot\psi_i)^2. \tag{13}$$

If we order the eigenvalues so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq \cdots$, we have

$$\begin{aligned} N\cdot E_\phi\{r^2\} &= \sum_i \lambda_i(h\cdot\psi_i)^2 \\ &= \sum_i \lambda_i\omega_i^2 \leq \lambda_1\sum_i \omega_i^2 = \lambda_1 \end{aligned} \tag{14}$$

and the maximum value is achieved when the filter function is taken to be the dominant eigenvector.

2.3. Signal to noise ratio and classification error

Several performance metrics are available for correlation filters that describe attributes of the correlation plane. The signal to noise ratio (SNR) is just one of them. Other useful quantities are the peak-to-correlation energy, the location of the correlation peak and the invariance to distortion. As correlation is typically used to locate and discriminate objects, another important measure of a filter's performance is how well it discriminates between different classes of objects. The simplest case is given by the discrimination between the signal and the noise. In this section we will show[14,22] that for the classical matched filter maximizing the SNR is equivalent to minimizing the probability of classification error P_e when the underlying probability distribution functions (PDFs) are Gaussians.

The classifier which minimizes the probability of error is the Bayes classifier. If we consider two normal distributions A and B, according to the Bayes decision rule, the observed vector $\boldsymbol{x} \in A$ if

$$\begin{aligned} &(\boldsymbol{x}-\boldsymbol{m}_A)^{\mathrm{T}}\Sigma_A^{-1}(\boldsymbol{x}-\boldsymbol{m}_A) - (\boldsymbol{x}-\boldsymbol{m}_B)^{\mathrm{T}}\Sigma_B^{-1}(\boldsymbol{x}-\boldsymbol{m}_B) \\ &\quad + \ln\frac{|\Sigma_A|}{|\Sigma_B|} < 2\ln\frac{P_A}{P_B} \end{aligned} \tag{15}$$

and $\boldsymbol{x} \in B$ otherwise, where $\boldsymbol{m}_A$, $\boldsymbol{m}_B$ are the distribution means, Σ_A, Σ_B the covariance matrices and P_A, P_B the occurrence probabilities.

Let us consider two classes: a deterministic signal corrupted with white Gaussian noise as class A and the noise itself as class B. In this case $\boldsymbol{m}_A = \phi$, $\boldsymbol{m}_B = \boldsymbol{0}$ and $\Sigma_A = \Sigma_B = I$. This means that the compo-

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.

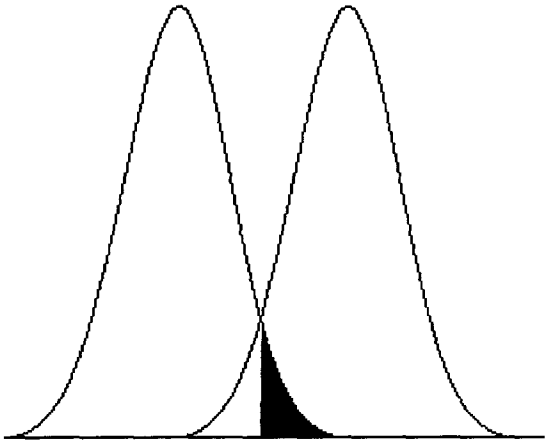

Fig. 1. The probability of error, represented by the shaded area, when the distributions are Gaussian with the same covariance.

nents of the signal are uncorrelated and have unit variance. If we further assume that the *a priori* probabilities of occurrence of these classes are equal, the probability of error (see also Fig. 1) is given by

$$P_e = \frac{1}{\sqrt{2\pi}} \int_{\eta}^{\infty} \exp(-u^2/2)\, du, \tag{16}$$

where $\eta = \frac{1}{2}\xi^{1/2}$, with ξ being the Mahalanobis distance between the PDFs of the two classes:

$$\xi = (\boldsymbol{m}_A - \boldsymbol{m}_B)^{\mathrm{T}} I (\boldsymbol{m}_A - \boldsymbol{m}_B) = \phi^{\mathrm{T}} \phi \tag{17}$$

and the Bayes decision rule simplifies to

$$\boldsymbol{x} \in A \text{ if } \phi^{\mathrm{T}} \boldsymbol{x} > \frac{1}{2}\xi, \tag{18}$$

$$\boldsymbol{x} \in B \text{ if } \phi^{\mathrm{T}} \boldsymbol{x} \leq \frac{1}{2}\xi. \tag{19}$$

The input vector $\boldsymbol{x}$ is then classified as signal or noise depending on the value of the correlation with the uncorrupted signal. We have already shown that correlation with the signal maximizes the signal to noise ratio, so when the noise distribution is Gaussian maximizing the SNR is equivalent to minimizing the classification error probability. When the noise is not white, the signal can be transformed by applying a whitening transformation A:

$$A^{\mathrm{T}} \Sigma A = I \tag{20}$$

and the previous reasoning can be applied.

3. SYNTHETIC DISCRIMINANT FUNCTIONS

While correlators are optimal for the recognition of patterns in the presence of white noise they have three major limitations: the output of the correlation peak degrades rapidly with geometric image distortions, the peak is often broad (see Fig. 2), making its detection difficult, and they cannot be used for multiclass pattern recognition. It has been noted that one can obtain better performance from a multiple correlator (i.e. one computing the correlation with several templates) by forming a linear combination of the resulting outputs instead of, for example, taking the maximum value.[23,24] The filter synthesis technique known as Synthetic Discriminant Functions (SDF) starts from this observation and builds a filter as a linear combination of MSFs for different patterns.[3,4] The coefficients of the linear combination are chosen to satisfy a set of constraints on the filter output, requiring a given value for each of the patterns used in the filter synthesis. By forcing the filter output to different values for different patterns, multiclass pattern recognition can be achieved. Let $\{\phi_i(x)\}_{i=1,\ldots,n}$ be a set of (linearly independent) images and $\boldsymbol{u} = \{u_1, \ldots, u_n\}^{\mathrm{T}}$ be a vector representing the required output of the filter for each of the images:

$$\phi_i \otimes h = u_i \tag{21}$$

where $\otimes$ represents correlation (not convolution). The filter h can be expressed as a linear combination of the images ϕ_i:

$$h(x) = \sum_{i=1,\ldots,n} b_i \phi_i(x) \tag{22}$$

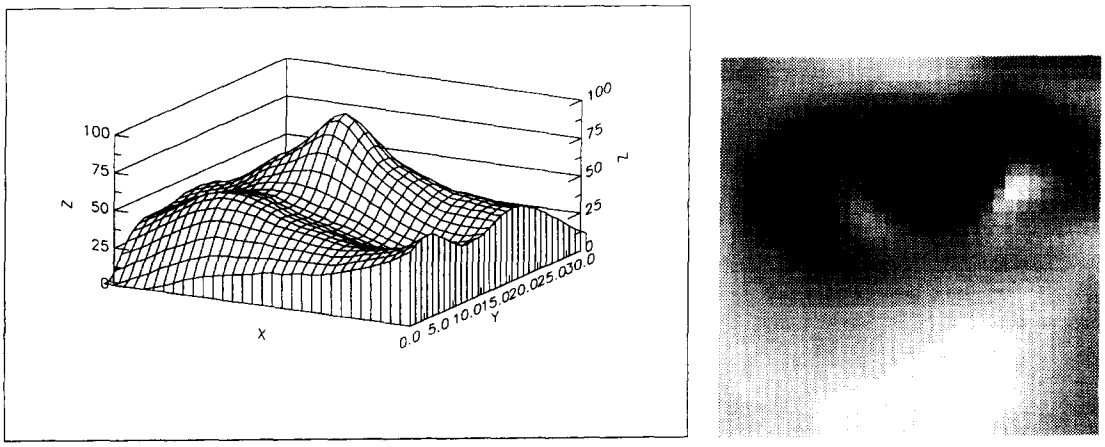

Fig. 2. The cross-correlation of the template reported on the right. Note the diffuse shape of the peak that makes its localization difficult.

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.

as any additional contribution from the space orthogonal to the images would yield a zero contribution when correlating with the image set. If we denote by $\boldsymbol{X}$ the matrix whose columns represent the images (represented as vectors by concatenating their rows), by enforcing the constraints we obtain the following set of equations:

$$\boldsymbol{b} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{u}, \tag{23}$$

which can be solved as the images are linearly independent. The resulting filter is appropriate for pattern recognition applications in which the input object can be a member of several classes and different distorted versions of the same object (or different objects) can be expected within each class. If M is the number of classes, n_i the number of different pattern within each class i, N the overall number of patterns, M filters can be built by solving

$$\boldsymbol{b}_i = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\delta_i, \quad i = 1, \ldots, M, \tag{24}$$

where

$$\delta_{ik} = \begin{cases} 1 & \sum_{j=1}^{i-1} n_j < k < \sum_{j=1}^{i} n_j, \\ 0 & \text{otherwise}, \end{cases} \tag{25}$$

$k = 1, \ldots, N$ and image ϕ_k belongs to class i if $\delta_{ik} = 1$. Discrimination of different classes can be obtained also using a single filter and imposing different output values. However the performance of such a filter is expected to be inferior to that of a set of class specific filters due to the high number of constraints imposed on the filter outputs.[3] While this approach makes it easy to obtain predefined values on a given set of patterns it does not allow to control the off-peak filter response. This can prevent reliable classification when the number of constraints becomes large.

The effect of filter *clutter* can also appear in the construction of a filter giving a fixed response over a set of images belonging to the same class (the Equal Correlation Filter introduced in reference (3)).

In order to minimize this problem we propose a new variant of SDFs: least squares SDFs. These filters are computed using only a subset of the training images[1] and the coefficient of the linear combination is chosen to minimize the square error of the filter output on all of the available images. In this case the matrix $\boldsymbol{R} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is rectangular and the estimate of the $\boldsymbol{b}$ relies on the computation of the pseudoinverse of $\boldsymbol{R}$:

$$\boldsymbol{R}^{\dagger} = (\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R})^{-1}\boldsymbol{R}^{\mathrm{T}}. \tag{26}$$

The dimension of the matrix to be inverted is $n \times n$, where n represents the number of images used to build the filter and not the (greater) number of training images. By using a reduced number of building templates the problem of filter cluttering is reduced. A different use of least square estimation for filter synthesis can be found in reference (4) where it is coupled to Karhunen–Loeve expansion for the construction of correlation SDFs.

The results for a sample application are reported in Fig. 3. Note that by using a least square estimate a good performance can be achieved using a small number of templates. This has a major influence on the appearance of the resulting MSF as can be seen in Fig. 4.

Another variant is to use *symbolic encoded filters*.[3] In this case a set of k filters is built whose outputs are 0 or 1 and can be used to encode the different patterns using a binary code. In order to use the filter for classification, the outputs are thresholded and the resulting binary number is used to index the pattern class.

Synthesis of the MSF from a projection SDF algorithm can achieve distortion invariance and retain shift invariance. However, the resulting filter cannot prevent large sidelobe levels from occurring in the correlation plane for the case of false (or true) targets. The next section will detail the construction of filters which guarantee controlled sharp peaks and good noise immunity.

4. ADVANCED SDFs

The signal to noise ratio maximized by the MSF is limited to the correlation peak: it does not take into account the off-peak response and the resulting filters often exhibit a sustained response well apart from the location of the central peak. This effect is usually amplified in the case of SDF when many constraints are imposed on the filter output. In order to locate the correlation peak reliably, it should be very localized.[5] However, it can be expected that the greater the localization of the filter response (approaching a δ function) the more sensitive the filter to slight deviations from the patterns used in its synthesis. This suggests that the best response of the filter should not really be a function, but some shape, like a Gaussian, whose dispersion can be tuned to the characteristics of the pattern space. In this section we will review the synthesis of such filters in the frequency domain.[12]

Let us assume for the moment that there is no noise. The correlation of the ith pattern with the filter h is represented by

$$z_i(n) = \phi_i(n) \otimes h(n), \quad n = 0, \ldots, d-1, \tag{27}$$

where d is the dimension of the patterns. In the following, capital letters are used to denote the Fourier transformed quantities. The filter is also required to produce an output u_i for each training image:

$$z_i(0) = u_i, \tag{28}$$

which can be rewritten in the Fourier domain as

$$\boldsymbol{H}^{+}\boldsymbol{X} = d\boldsymbol{u}, \tag{29}$$

where $^{+}$ denotes complex conjugate transpose. Using Parseval's theorem, the energy of the ith *circulant* cor-

[1] The subset of training images can be chosen in a variety of ways. In the reported experiments they were chosen at random. Another possibility is that of clustering the available images, the number of clusters being equal to the number of images used in filter synthesis.

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.