Face Detection by Support Vector Machines in the Walsh Transform Domain*

A. Fazekas**, C. Kotropoulos, I. Buciu*** and I. Pitas

Department of Informatics, Aristotle University of Thessaloniki
GR-54006 Thessaloniki Box 451, Greece
fattila@math.klte.hu,{costas,nelu,pitas}@zeus.csd.auth.gr

Abstract. This paper reports first experimental results in order to access the properties of the support vector machines in the Walsh Transform Domain for face detection. We prove theoretical results about the VC-dimension of the support vector machines which are built in the space of the two-dimensional (2-D) Walsh functions. Morever, we demonstrate by experiments that support vector machines in the Walsh Transform Domain can separate more efficiently face patterns from non-face ones in the sense that the margin between the two classes of patterns is increased.

1 Introduction

The Bayes likelihood ratio test yields the optimal classifier in the sense that it minimizes the probability of error [4]. However in order to construct the likelihood ratio, the conditional probability density function (pdf) for each class must be known. Although, there are several procedures for estimating a pdf from a finite number of observations [4], the problem of density estimation is ill-posed [8]. An alternative method to solve a two-class pattern recognition problem is to resort to example-based techniques, such as the support vector machines [8].

Support vector machines implement the following idea: Let us map the input vectors, which are the elements of the training set, onto a high-dimensional feature space through a mapping chosen a priori. In this space, we construct the optimal separating hyperplane to get a binary decision whether the input vector belongs to the given class or not. For example, in face detection the input vector comprises the gray levels of pixels from a rectangular region of the digital image and the result of the binary decision is the answer whether this region is face or not.

In general the determination of the separating hyperplane is not easy, because the dimensionality of the feature space is high. However, in Hilbert spaces one can estimate the inner product of two vectors in the feature space as a function of two vectors

* This work was supported by the European Union Research Training Network "Multi-modal Human-Computer Interaction (HPRN-CT-2000-00111)".

** A. Fazekas is on leave from Institute of Mathematics and Informatics, University of Debrecen H-4010 Debrecen P.O.Box 12, Hungary

*** I. Buciu is on leave from University of Oradea, RO-3700 Oradea, Armatei Romane 5, Romania

in input space. These expressions for inner products are referred as *kernel functions*. Some kernel functions are well-known, for example the polynomial, the radial, the sigmoid, etc. [1],[2]. In this paper we will construct support vector machines in the Walsh Transform Domain. In this case the feature space where the separating hyperplane will be constructed, gives a good description of the special symmetries of the input vector, because the mapping of the input vector into the feature space is based on the 2-D Walsh transform. The 2-D *Walsh transform* is the sum of the values of the function, which are multiplied by either $+1$ or -1 depending on the co-ordinates of the points. This transform is useful in the pattern recognition [3].

The VC-dimension is the capacity factor of the support vector machine [9], so its knowledge is very important to control the behavior of the support vector machine. In this paper we are going to prove several propositions about the VC-dimension of the class of the 2-D Walsh functions.

The structure of the paper is as follows. Section 2 is a brief overview of the Walsh system. Section 3 explains the construction of the 2-D Walsh kernel. Section 4 describes the theoretical results on the VC-dimension of the class of the 2-D Walsh functions. Section 5 reports the first promising experimental results.

2 The Walsh System

In the literature the term "Walsh functions" refers to one of three orthonormal systems: the *Walsh-Paley system*, the *original Walsh system*, or the *Walsh-Kaczmarz system* [7]. These systems contain the same functions and differ only in enumeration. We will investigate the Walsh-Paley system which will be referred as the *Walsh system* henceforth. For more details the interested reader may consult [7].

Notation 1. *We will denote the set of non-negative integers by* $\mathbb{N}$*, the set of positive integers by* $\mathbb{N}^+$*, the set of integers by* $\mathbb{Z}$*, and the set of real numbers by* $\mathbb{R}$*.*

Definition 1. *Let* r *be the function defined on* $[0,1)$ *by*

$$r(x) = \begin{cases} 1, & x \in [0, \frac{1}{2}), \\ -1, & x \in [\frac{1}{2}, 1) \end{cases}$$

extended to $\mathbb{R}$ *by periodicity of period 1. The Rademacher system* $R = \{r_n(x), n \in \mathbb{N}\}$ *is defined by*

$$r_n(x) = r(2^n x), \qquad x \in \mathbb{R}, n \in \mathbb{N}.$$

Definition 2. *Given* $n \in \mathbb{N}$ *it is possible to write* n *uniquely as*

$$n = \sum_{k=0}^{\infty} n_k 2^k,$$

where either $n_k = 0$ *or* 1 *for* $k \in \mathbb{N}$*. This expression will be called the binary expansion of* n *and the numbers* n_k *will be called the binary coefficients of* n*.*

Definition 3. *Let* x *be an arbitrary element of the interval* $[0,1)$*. If* x *has the form* $\frac{p}{2^n}$ *for some* $p, n \in \mathbb{N}$ $(0 \leq p < 2^n)$*, we will call* x *dyadic rational in the interval* $[0,1)$*.*

Definition 4. *Any $x \in [0,1)$ can be written in the form*

$$x = \sum_{k=0}^{\infty} x_k 2^{-(k+1)},$$

where each x_k is equal to either 0 or 1. We will call it the dyadic expansion of x. When x is a dyadic rational there are two expressions of this form, one which terminates in 0's and one which terminates in 1's. In this case, the dyadic expansion of x we will mean the one which terminates in 0's.

Definition 5. *The Walsh system $W = \{w_n(x), n \in \mathbb{N}\}$ is product of Rademacher functions in the following way. If $n \in \mathbb{N}$ has binary coefficients $\{n_k, k \in \mathbb{N}\}$ then*

$$w_n(x) = \prod_{k=0}^{\infty} r_k^{n_k}(x).$$

It is easy to see that this product is always finite, $w_0 = 1$ and $w_{2^n} = r_n$ for $n \in \mathbb{N}$. It is worth noticing that each Walsh function is piecewise constant with finitely many jump discontinuities on $[0,1)$, and takes only the values of either $+1$ or -1.

Definition 6. *The 2-D Walsh system $W^{(2)} = \{w_{(n,m)}(x,y) \mid n,m \in \mathbb{N}\}$ is defined in the following way:*

$$w_{(n,m)}(x,y) = w_n(x) \cdot w_m(y).$$

3 The Construction of the Walsh Kernel

In this section we define a new kernel function for support vector machines. The construction is based on Vapnik's idea [9].

It is well-known, that the Walsh system is a complete orthonormal system on $[0,1)$ and the Walsh system is a Schauder basis in L^p for $1 < p < \infty$ [7]. This linear space is a Hilbert space, where the inner product, which is denoted by $\langle \cdot, \cdot \rangle$, is the integral of the product of two functions [7].

It is well-known that if f is an $\mathbb{R}$-valued, integrable function on the interval $[0,1)$ then

$$f(x) = \sum_{k=0}^{\infty} a_k(f) w_k(x), \tag{1}$$

where $a_k = \langle f, w_k \rangle$.

In the case of the 2-D Walsh system (1) takes the following form:

$$f(x,y) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a_{(i,j)}(f) w_{(i,j)}(x,y).$$

Let N and M be non-negative integers, and

$$\mathbf{\Phi}_{(N,M)}(f) = \left(a_{(0,0)}(f), \ldots, a_{(0,M-1)}(f), \ldots, a_{(N-1,M-1)}(f)\right).$$

We would like to note it is convenient to assume that $N = 2^n$, $M = 2^m$ and $n, m \geq 2$, where $n, m \in \mathbb{N}$. The reason for these relations stems from the theoretical characteristics of the Walsh functions [7].

Let f and g be $\mathbb{R}$-valued, integrable functions on the interval $[0,1)^2$. Then the kernel function is

$$K_{(N,M)}(f,g) = \left\langle \mathbf{\Phi}_{(N,M)}(f), \mathbf{\Phi}_{(N,M)}(g) \right\rangle .$$

It is easy to see that $K_{(N,M)}(f,g)$ is simply an ordinary inner product in $N \cdot M$-dimensional feature space, where $\mathbf{\Phi}_{(N,M)}(f)$ is a Walsh transform of the two-dimensional function f. This transform is very informative, because it is the sum of the values of the function, i.e. the gray level values, which are multiplied by either $+1$ or -1 depending on the co-ordinates of the pixels. In digital image processing the functions f and g are finite functions, whose domains are sets of connected finite subsets of $\mathbb{Z}^2$. In this case, the meaning of the function $\mathbf{\Phi}$ is the discrete 2-D Walsh transform of the function f [5].

For example, let us consider the first elements of the vector $\mathbf{\Phi}(f)$. The first element is the sum of the values of the function in the points of the finite subset, the second element is the difference between the sum of the values of the function on the first half part and the sum of the values of the function on the second half part. If the second element of the vector is equal to 0, then there is a balance between the avarage values of the function on these parts. So the elements of this vector are special measures of the "symmetries" of the function, which can be useful for describing the image [3].

To determine the value of the kernel function $K_{(N,M)}(f,g)$ is not a difficult task, because fast Walsh transforms are well-known in the literature [5].

4 The VC-dimension of the Class of the 2-D Walsh Functions

The Vapnik-Chervonenkis dimension has a very important role in the statistical learning. The VC-dimension of the support vector machines characterizes the learning capacity of the machine. With control of the VC-dimension one can avoid overfitting of the support vector machines and one can minimize the expected value of the error [9]. So the knowledge of the VC-dimension of the class of the functions employed in the learning algorithm is very important.

At first we quote some definitions from [9], which are important to understand the theoretical results of this section.

Definition 7. *An arbitrary $\{+1,-1\}$-valued function with domain $\mathbb{R}^2$ is called an 2-D indicator function.*

Definition 8. *Let f be an arbitrary 2-D indicator function. The sets $\{(x,y) \mid f(x,y) = +1,\ x,y \in \mathbb{R}\}$ and $\{(x,y) \mid f(x,y) = -1,\ x,y \in \mathbb{R}\}$ are the separated classes of the domain by using f.*

Definition 9. *The VC-dimension of a set of 2-D indicator functions is equal to the largest number h of points of the domain of the functions that can be separated into two different classes in all the 2^h possible ways using a function from this set. If for any n there exists a set of n points that can be shattered by the functions of the set, then the VC-dimension is equal to infinity.*

In the rest of the paper we assume that the domain of 2-D Walsh functions and of the 2-D indicator functions is the set $[0,1)^2$.

Lemma 1. *Let x and y be arbitrary elements of the interval $[0,1)$. Let $x_0, x_1, \ldots$ and $y_0, y_1, \ldots$ be the dyadic expansions of x and y, respectively. Let n and m be arbitrary elements of $\mathbb{N}$, and $n_0, n_1, \ldots$ and $m_0, m_1, \ldots$ be the binary coefficients of n and m, respectively. Then the following equation holds:*

$$w_{(n,m)}(x,y) = (-1)^{\sum_{k=0}^{\infty} n_k x_k} \cdot (-1)^{\sum_{k=0}^{\infty} m_k y_k}.$$

Proof. By using Definitions 5 and 6. □

Theorem 1. *The VC-dimension of the class of all the 2-D Walsh functions is equal to ∞.*

Proof. From Definition 6 one can see that the 2-D Walsh system is simply a set of the products of two one-dimensional (1-D) Walsh functions. Therefore, it suffices to prove the theorem for the 1-D case, or equivalently for points $(x_i, 0)$, $i = 0, \ldots, h-1$.

The proof will be constructive. Let h be arbitrary elements of the set $\mathbb{N}^+$. Let $\mathbf{x}_0, \ldots, \mathbf{x}_{2^h-1}$ be all the h-dimensional vectors with elements 0 and 1. Let us construct a matrix in the following form:

$$X = \left(\mathbf{x}_0, \ldots, \mathbf{x}_{2^h-1} \right).$$

Let us consider the rows of the matrix X as the dyadic expansion of h numbers in the following form: The dyadic expansion of the ith $(0 \le i \le h-1)$ number is $x_{i0}, \ldots, x_{i2^h-1}, 0, \ldots$. We will denote these numbers by p_i.

Let $v_0, \ldots, v_{h-1}$ be an arbitrary sequence of the values $+1$ and -1. We will show that one from all the 2-D Walsh functions admits the values $v_0, \ldots, v_{h-1}$ at the points $(p_0, 0), \ldots, (p_{h-1}, 0)$ of the set $[0,1)^2$.

Let us select an index t, which fulfils the following condition

$$x_{it} = \begin{cases} 0, \text{ if } v_i = 1, \\ 1, \text{ if } v_i = -1. \end{cases}$$

Let n be equal to 2^t. By Lemma 1, $w_{(n,m)}(p_i, 0) = (-1)^{\sum_{k=0}^{\infty} n_k x_{ik}} \cdot (-1)^{\sum_{k=0}^{\infty} m_k 0_k}$. Because $n = 2^t$, only n_t is equal to 1, and the other binary coefficients of the n are equal to 0. On the other hand, because y is equal to 0, so the dyadic expansion of y contains only 0's. From this follows that $w_{(n,m)}(p_i, 0) = (-1)^{x_{it}} = v_i$. So, the 2-D Walsh functions can shatter these points. Because the number h is arbitrary, so the VC-dimension of the set of all 2-D Walsh functions is equal to ∞. □

The relations $N = 2^n$, $M = 2^m$ and $n, m \ge 2$ are assumed in the following.

Theorem 2. *The VC-dimension of the set $W_{(N,M)} = \{w_{(k_1,k_2)}(x,y) \mid k_1 = 0, \ldots, N-1, k_2 = 0, \ldots, M-1\}$ is equal to $\log_2(N \cdot M)$.*

Proof. At first, we will prove the VC-dimension of the set $W_{(N,M)}$ is not less than $\log_2(N \cdot M) = n + m$. By Definition 9 it is enough to find $n + m$ points in the set $[0,1)^2$, which can be shattered by the elements from $W_{(N,M)}$.

Let us consider $n+m$ points in the set $[0,1)^2$ in the following form: $(\frac{1}{2^1}, 0), \dots, (\frac{1}{2^n}, 0), (0, \frac{1}{2^1}), \dots, (0, \frac{1}{2^m})$. Let us call them as p_i of co-ordinates (x_i, y_i) $(0 \le i \le n+m-1)$. Let $l_0, \dots, l_{n+m-1}$ be an arbitrary sequence of values $+1$ and -1. We show that one from all 2-D Walsh functions in $W_{(N,M)}$ admits the values $l_0, \dots, l_{n+m-1}$ at the points $p_0, \dots, p_{n+m-1}$.

Let $u_0, \dots, u_{n-1}, 0, \dots$ be the binary expansion of u, and $v_0, \dots, v_{m-1}, 0, \dots$ be the binary expansion of v, respectively. Let us define the binary expansion of u and v as follows.

$$u_i = \begin{cases} 0, \text{if } l_i = 1, \\ 1, \text{if } l_i = -1, \end{cases} (0 \le i \le n-1), \qquad v_{i-n} = \begin{cases} 0, \text{if } l_i = 1, \\ 1, \text{if } l_i = -1, \end{cases} (n \le i \le n+m-1).$$

Then the function $w_{(u,v)}(x,y) \in W_{(N,M)}$ admits l_i value at the point p_i of co-ordinates (x_i, y_i), because by Lemma 1

$$w_{(u,v)}(x,y) = (-1)^{\sum_{k=0}^{\infty} u_k x_k} \cdot (-1)^{\sum_{k=0}^{\infty} v_k y_k}.$$

From this follows

$$w_{(u,v)}(x_i, y_i) = \begin{cases} (-1)^{u_i} = l_i, \quad \text{if } 0 \le i \le n-1, \\ (-1)^{v_{i-n}} = l_i, \text{if } n \le i \le n+m-1. \end{cases}$$

On the other hand, let us consider $\log_2(N \cdot M) + 1$ points in the set $[0,1)^2$. Let us suppose we can find functions $w(x,y)$ from the set $W_{(N,M)}$, which can shatter these points.

It is very easy to check that the number of elements of $W_{(N,M)}$ is $N \cdot M$. The number of all vectors of size $\log_2(N \cdot M) + 1$ is equal to $2^{\log_2(N \cdot M)+1} = N \cdot M \cdot 2$, but there are only $N \cdot M$ functions, so the elements of the set $W_{(N,M)}$ can not shatter these points. So the VC-dimension of the set $W_{(N,M)}$ is not greater than $\log_2(N \cdot M)$. □

Definition 10. *Let $G = \{x_n \mid n \in \mathbb{N}\}$, where either $x_n = 0$ or $x_n = 1$. Set $I_0(x) = G$ for all $x \in G$. For each $x \in G$ and $n \in \mathbb{N}^+$ define*

$$I_n(x) = \{y \in G \mid y_i = x_i, 0 \le i < n\}.$$

We call the sets $I_n(x)$ the dyadic intervals of order n in $[0,1)$.

Definition 11. *By a dyadic step function of order n we mean a finite linear combination of characteristic functions of dyadic intervals of order n in $[0,1)$.*

Definition 12. *By a 2-D dyadic step function of order (n,m) we mean a product of two dyadic step functions of orders n and m, respectively.*

Notation 2. *We use the following notation:*

$$\theta(x) = \begin{cases} 1, \quad \textit{if } x >= 0, \\ -1, \textit{if } x < 0, \end{cases}$$

where $x \in \mathbb{R}$.

Definition 13. *Let $f_k(x,y)$ be $\mathbb{R}$-valued functions. We call the set of indicator functions $\theta(f_k(x,y)-t)$, where $t \in \left(\inf_{(x,y)} f_k(x,y), \sup_{(x,y)} f_k(x,y)\right)$, the set of indicators for functions $f_k(x,y)$.*

Theorem 3. *The VC-dimension of the set $lin(W_{(N,M)}) = \{w(x,y) \mid w(x,y) = \alpha_{(0,0)} w_{(0,0)}(x,y) + \ldots + \alpha_{(0,M-1)} w_{(0,M-1)}(x,y) + \ldots + \alpha_{(N-1,M-1)} w_{(N-1,M-1)}, \alpha_{(i,j)} \in \mathbb{R}\}$ is equal to $N \cdot M$.*

Proof. Let f be an arbitrary finite linear combination of the elements in $W_{(N,M)}$. By Definition 5, the function f takes the following form:

$$f(x,y) = \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} \alpha_{(i,j)} w_i(x) \cdot w_j(y) = \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} \alpha_{(i,j)} \prod_{k=0}^{n} r_k^{i_k}(x) \prod_{k=0}^{m} r_k^{j_k}(y),$$

where $\alpha_{(i,j)} \in \mathbb{R}$. So the function $f(x,y)$ is a finite linear combination of finite products of the first $\max(n,m)$ functions in the Rademacher system. The elements of Rademacher system are piecewise constant with finitely many jump discontinuities on $[0,1)$. By Definition 1, $r_n(x)$ has $2^n - 1$ jumps, so the function $f(x,y)$ divides up the set $[0,1)^2$ into $N \cdot M$ parts, at most, by the continuity of the finite linear combination of continuous functions. Let us suppose that the function $f(x,y)$ divides up the set $[0,1)^2$ into $N \cdot M$ parts.

By [7] it is easy to prove that any 2-D dyadic step function of order (n,m) is a finite linear combination of the elements of the set $W_{(N,M)}$. By Definition 13 it is enough to investigate the 2-D dyadic step functions of order (n,m), which are indicator functions.

Let $x_0, \ldots, x_{N \cdot M-1}$ be elements of the $N \cdot M$ parts. Then these points can be shattered by the elements in $lin(W_{(N,M)})$ by the previously established connection between the 2-D dyadic step functions and the finite linear combinations of the elements in $W_{(N,M)}$. So the VC-dimension is not less than $N \cdot M$.

Let us suppose the VC-dimension is greater than $N \cdot M$. Let $x_0, x_1, \ldots, x_{N \cdot M}$ be the points in the set $[0,1)^2$, which can be shattered by the finite linear combinations of the elements in $W_{(N,M)}$. But in this case, two points of them are in the same part, independ of the selection of the function in $lin(W_{(N,M)})$, by the structure of the Rademacher functions. So every function in $lin(W_{(N,M)})$ admits the same values, which is a contradiction. □

Corollary 1. *The VC-dimension of the kernel function $K_{(N,M)}$ is equal to $N \cdot M$.*

5 Experimental Results

For all experiments the `Mathlab` SVM toolbox developed by Steve Gunn was used [6]. For a complete test, several auxiliary routines have been added to the original toolbox.

A training data set of 46 images, 31 images containing a face and another 15 images with non-face patterns, is built. The images containing face patterns have been derived from the face database of IBERMATICA where several sources of degradation are modeled. For a description of this database the interesting reader may refer to [10].

All images in this database are recorded in 256 grey levels and they are of dimensions 320×240. The procedure for collecting face patterns is as follows. From each image a bounding rectangle of dimensions 128×128 pixels has been manually determined that includes the actual face. This area has been subsampled four times. At each subsampling, non-overlapping regions of 2×2 pixels are replaced by their average. Accordingly, training patterns of dimensions 8×8 are built. The ground truth, that is, the class label $+1$ has been appended to each pattern. Similarly, 15 non-face patterns have been collected from images in the same way, and labeled by -1.

We have trained the three different SVMs indicated in Table 1. The trained SVMs have been applied to 414 test images (249 face and 165 non-face) from the IBERMATICA database that have not been included in the training set. The resolution of each test image has been reduced four times yielding a final image of dimensions 8×8. The test images are classified as either non-face or face ones.

Table 1 summarizes the results of the experiments.

Table 1. Experimental Results

	Linear	Walsh	Polynomial
Time (sec)	2.73	2.65	2.92
Number of Errors	10	9	8
Margin	0.6	4.2	3.4
Number of SVs	12	7	9

The first row in Table 1 depicts the time needed for test experiments using each kernel functions. As can be seen, the support vector machines that are based on the Walsh functions require less time than the other SVMs. The numbers of errors shown in the second row are the misclassification errors. That is the number of real faces either classified as non-faces, or non-face instances classified as faces. From this point of view the Walsh kernel is in the middle. Perhaps the most significant improvement is that the margin that separates the two classes admits the largest value in the case of Walsh kernel. Finally, the number of support vectors in the case of Walsh kernel lies in between the number of support vectors for the linear and the polynomial case. We conclude that the performance of support vector machines based on the Walsh kernel function is in between the performance of the linear and the polynomial support vector machines.

Fig. 1. Example of false detection. The polynomial SVM detected this image as a face.

Fig. 2. Example of good detection.

References

1. C.J.C. Burges. A tutorial on support vector machines for pattern recognition, *Knowledge Discovery and Data Mining*, **2**, 1998.
2. L. Devroye, L. Györfi and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.
3. A. Fazekas and A. Hajdu. Recognizing typeset documents using Walsh transformation, *Journal of Computing and Information Technology*, *Accepted for publication*, 2000.
4. K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, 1990.
5. R.C. Gonzalez and R.E. Woods. *Digital Image Processing*. Addison-Wesley, New York, 1993.
6. S. Gunn. Support vector machines for classification and regression, *ISIS Technical Report ISIS-1-98*, Image Speech Intelligent System Research Group, University of Southapton, 1998.
7. F. Schipp, W.R. Wade, P. Simon and J. Pál. *Walsh Series: An Introduction to Dyadic Harmonic Analysis*. Adam Hilger, Budapest, 1990.
8. V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.
9. V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.
10. C. Kotropoulos, A. Tefas, and I. Pitas. Morphological elastic graph matching applied to frontal face authentication under well-controlled and real conditions, *Pattern Recognition* **33**, 31-43, 2000.