International Journal of Information Acquisition

Margin and Domain Integrated Classification for Images

Yen-Lun Chen
Center for Intelligent and Biomimetic Systems
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
Shenzhen 518055, China.

Yuan F. Zheng
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, OH 43210, USA.

Yi Liu
PIPS Technology, A Federal Signal Company
Knoxville, TN 37932, USA.

Received (to be inserted by publisher)

Multi-category classification is an on going research topic in image acquisition and processing for numerous applications. In this paper, a novel approach called margin and domain integrated classifier (MDIC) is addressed. It merges the conventional support vector machine (SVM) and support vector domain description (SVDD) classifiers, and handles multi-class problems as a combination of several target classes plus outliers. The basic idea behind the proposed approach is that target classes possess structured characteristics while outliers scatter around in the feature space. In our approach the domain description and large-margin discrimination are adjustable and therefore yield higher classification accuracy which leads to better performance than conventional classifiers. The properties of MDIC are analyzed and the performance comparisons using synthetic and real data are presented.

Keywords: Margin and domain integrated classification (MDIC), pattern classification, multi-category classification, support vector machine (SVM), support vector domain description (SVDD).

1. Introduction

Pattern classification [Duda *et al.*, 2001] has been an essential subject in machine learning, computer vision, image and video processing for a long time. It is widely used in numerous applications including handwritten digit recognition, 3-D object recognition [Pontil & Verri, 1998], object detection [Kim & Jo, 2009], human face detection [Osuna *et al.*, 1997], and image annotation [Goh *et al.*, 2005]. In recent years, emerging applications such as video object extraction, content-based image retrieval, biometric data verification, and medical image diagnosis, etc., demand even more powerful tools. For

example of detecting a small and particular set of targets in a surveillance video around urban area, the appearance of targets could vary significantly which are surrounded by various kinds of ordinary objects such as buildings of schools and hospitals and civilian vehicles. Automatic detection of targets of interests in such an environment is extremely difficult and challenging.

Many approaches have been proposed for object recognition, which in most cases is achieved by making use of multiple samples of the object in establishing the model. Among all the approaches, statistical model training [Hastie *et al.*, 2001] has attracted a great deal of attention in recent years. More recently, large margin approaches, particularly support vector machine (SVM), have become a popular tool for object classification [Cortes & Vapnik, 1995][Lin & Wang, 2002][Platt, 1999][Vapnik, 1999]. SVM stresses maximal margin-discrimination between two or multiple classes, while another model training approach called support vector domain description (SVDD) emphasizes domain-description [Tax & Duin, 1999] [Scholkopf *et al.*, 2001]. When each approach is applied in reality, either the discrimination or the description approach has inherited disadvantages which cannot meet the needs of emerging applications as mentioned earlier.

Domain-description methods such as SVDD focus on estimating the distribution of the regular (or non-outlier) portion of the data. The problem of domain-description is a special type of classification problem. For example in one-class classification, we are always dealing with a two-class classification problem, where the two classes are called the target and the outlier class respectively. The concept was developed under the assumption that only samples for the target classes are available, and no non-target class samples are used for training. In reality, outlier class can be sampled very sparsely, or can be totally absent for the reason that it might be very expensive or difficult to do the measurements on these types of objects. Another extreme case is when the outliers are so abundant that a good sampling of the outliers is not possible [Tax, 2009]. On the other hand, margin-discrimination approaches such as SVM focus on the optimal separation between different classes from the knowledge of training samples. The approach was developed under the assumption of large amount of training samples (including sufficient outliers) available for learning the model parameters. However, the intrinsic nature of outliers is that they are scarce, unpredictable, and distant from others, so no training data set can possibly contain all forms of outliers [Goh *et al.*, 2005]. As a result, the situation of object classification in reality is usually to minimize a prediction risk based on limited training samples.

To be more specific, description-based approach describes a closed subspace to accommodate the samples. The domain of the subspace is minimized through training samples, which is therefore tight and no separation occurs. As a result, any sample that is nearby the domain but falls out is not identified as a member of the class. In other words, training results of the description-based approach are over-fitting to an object class. The over-fitting problem causes an elevated *miss* rate in many emerging applications such as face recognition for terrorist identification. Discrimination-based approaches on the other hand are optimized for a large margin between different classes, but do not evaluate the "description" of the sample for its fitness to a particular class. In other words, training results of the discrimination-based approach are under-fitting to an object class. Using the discrimination-based approach, one sample is classified to one particular class, which, however, may belong to neither of the classes. The problem of unable to reject an uncharacteristic sample causes an elevated *false-alarm* rate in several practical applications [Yuan & Casasent, 2003].

The above two problems are further illustrated in Fig. 1, where the vertical and the horizontal axes represent two different features of an object. In (a), the description-based approach generates the tightest hyperspheres for describing each class, and in (b), the discrimination-based approach generates hyperplanes with large margins to separate the classes. When each approach is applied alone, (c) shows that either outliers are classified to the target class as false alarms by the discrimination-based approach or

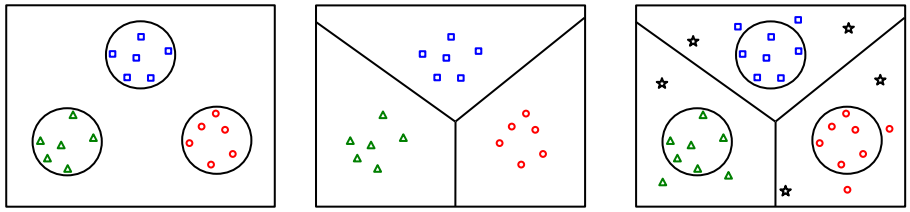


Fig. 1. Description- or discrimination-based approach generates large classification errors in application: (a) the description-based approach creates a domain for each class, (b) the discrimination-based approach separates the space into multiple classes, and (c) both approaches generate errors where target samples (triangles, circles, or squares) are missed by the description-based approach or non-target samples (shown as stars) are classified as target by the discrimination-based approach.

target samples are missed by the description-based approach.

As in the foregoing discussion, existing approaches rely on either discrimination or description through training samples. By far, SVM has been the major tool for discrimination classification and received a great deal of attention in recent years. SVM was originally developed to differentiate two as opposed to multiple classes of objects [Cortes & Vapnik, 1995]. In reality, many applications require a machine to discriminate multiple classes of objects. To treat multi-object classification, approaches have been developed using a generalized approach of two-class SVM by [Liu & Zheng, 2005] and [Liu & Zheng, 2008]. In the literature, more relevant works can be found in the field of SVM for both theory and applications. There the emphasis has always been in discrimination, that is, maximizing the separation margin between two or multiple classes which is the primary goal of the generalized SVM. In the context, however, the margin based approach has its limitation that is to misclassify non-target samples. To overcome this practical problem, the description-based approach was invented in 1999 extending the concept of support vectors from SVM for one-class data or domain description [Tax & Duin, 1999] [Scholkopf *et al.*, 2001]. Just as SVM, SVDD continues to identify applications in object classification such as facial expression analysis [Zeng *et al.*, 2006], image classification and retrieval [Lai *et al.*, 2004], hyperspectral anomaly detection [Banerjee *et al.*, 2007], and many others [Li & Hao, 2008]. SVDD has along with SVM become a popular tool for object classification. In reality, the one-sample approach can be considered as an exception of SVDD in which a single sample serves as the center of a domain, and the radius of the domain is determined by a threshold selected according to the nature of application. In SVDD, however, both are determined by the support vectors through an optimization process for achieving better performance statistically due to the use of training samples. Again, SVDD is limited to applications in which samples are closely clustered in a space. Even minor variation of the object may result in misclassification.

To our knowledge, there does not exist any work that *systematically* integrates the two advantages of *margin* for discrimination and *domain* for description together. So far only several works have mentioned the idea of margins in the framework of one-class classification or description. Yuan and Casasent developed an approach constructing multiple one-class SVMs with one for each class [Yuan & Casasent, 2003]. If all the classifiers reject an input, the sample is not classified to any target class; otherwise, it is accepted as the class with the highest level of confidence. More recently, Liu and Zheng developed a binary classifier called minimum enclosing and maximum excluding machine (MEMEM) which takes outliers into consideration [Liu & Zheng, 2006]. Similar to SVDD, MEMEM models the support of a target class by a hypersphere, but unlike SVDD it seeks an additional hypersphere that excludes the negative samples by a wide shell. To integrate margin and domain in a sin-

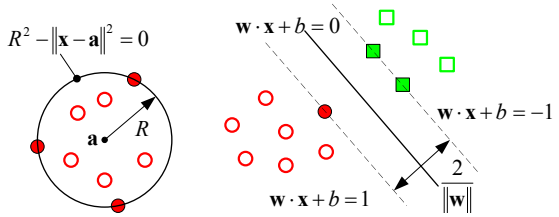
gle model for multiple classification, we extend the capabilities of MEMEM to a much powerful tool, which is called margin and domain integrated classification (MDIC) in the reminder of this paper. MDIC integrates the notion of margin into the description-based approach for further enhancing its performance. It makes both the domain and the separation margin flexible so that the optimal performance can be realized to adapt various purposes and conditions of applications. By doing so, the discriminating ability of the classifier is enhanced while its descriptive ability is preserved. The most important difference between MDIC and MEMEM is that the formulation of MEMEM is for the binary case, while MDIC is for multiple classes. That is, the formulation of MDIC is a generalization from MEMEM, which represents a significant improvement from MEMEM. As a result, the applications of MDIC is greatly expanded from two classes to generally-multiple classes, while MEMEM is limited to one target class and one outlier class.

It is still an ongoing research topic to extend a classifier from binary to multiple categories [Weston & Watkins, 1999][Hsu & Lin, 2002][Lee *et al.*, 2004][Rifkin & Klautau, 2004] [Liu *et al.*, 2005][Joshi *et al.*, 2010]. Conventionally, a single multi-class problem is considered as a collection of multiple binary problems. In one-versus-all (OVA) method, k classifiers are constructed for each class to separate it from the rest of the classes. In one-versus-one (OVO) method, $k(k-1)/2$ pairs of classifiers are constructed to separate each class from another one, and the decision function is determined by combining the results. OVA approach has been widely used in the literature to solve multi-class problems. However, as pointed out in [Lee *et al.*, 2004], OVA approach performs poorly when there does not exist a dominating class with conditional probability greater than 0.5. Therefore, a true extension from binary to multi-category classification which considers all classes simultaneously is desirable. We consider a multi-class classification problem in the framework of both discrimination and description. From this perspective, margin and domain integrated classification (MDIC) enables a classifier to possess both margin-discrimination and domain-description capabilities such that it can be more powerful and robust than existing classifiers in many emerging applications. Furthermore, it treats multi-class classification in parallel, rather than sequentially, to achieve an optimal performance. Ultimately it is superior to any method which is based on either discrimination or description only.

The rest of this paper is organized as follows. In Section 2, we begin with a review of the related work on binary learning machines. Then we present the mathematical formulation of MDIC method to illustrate the mechanism for integrating discrimination and description in Section 3. Experimental results are provided in Section 4 which is followed by conclusions in Section 5.

2. Binary Learning Machines

Before presenting detailed formulations of MDIC in the next section, we first summarize the mathematical formulations of binary classifiers: support vector domain description (SVDD) and support vector machine (SVM) in this section. The goal of binary classification is to find a boundary in vector space R^d to separate two different classes labeled as $\{+1, -1\}$, from the knowledge of a training set which contains n samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i \in R^d$. In the field of optimization, this problem is mathematically modeled as a constrained minimization, which can be transformed from primal to dual by the Lagrange method, and then solved via quadratic programming. At the end of this section, comparisons between SVDD and SVM are summarized in Table 1.



(a)SVDD (b)SVM

Fig. 2. Binary-class learning machines: (a) support vector domain description, (b) support vector machine.

2.1. *SVDD*

Support vector domain description (SVDD), proposed by Tax and Duin, seeks the minimum hypersphere that encloses all of the training samples labeled as class +1 [Tax & Duin, 1999]. As illustrated in Fig. 2(a), SVDD defines a hypersphere $B(\mathbf{a}, R)$, described by center $\mathbf{a}$ and radius R, with minimum volume to contain all (or most of) the data objects $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ as the domain of a class. Mathematically, the problem is written as

$$\begin{aligned} &\min : R^2 + C\sum_i \xi_i, \\ &\text{subject to} : \begin{cases} ||\mathbf{x}_i - \mathbf{a}||^2 \le R^2 + \xi_i, \\ \xi_i \ge 0, \end{cases} \end{aligned} \tag{1}$$

where slack variables ξ_i are introduced in the minimization to accommodate potential training errors and C is a parameter to adjust the weight of training errors.

Using Lagrange multipliers α_i, the above primal problem (1) is transformed to the following dual problem,

$$\begin{aligned} &\min : \sum_i \sum_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j - \sum_i \alpha_i \mathbf{x}_i \cdot \mathbf{x}_i, \\ &\text{subject to} : \begin{cases} \sum_i \alpha_i = 1, \\ 0 \le \alpha_i \le C. \end{cases} \end{aligned} \tag{2}$$

The optimization of (2) is a quadratic programming problem, by solving which we obtain the center of hypersphere to be a linear combination of data objects: $\mathbf{a} = \sum_i \alpha_i \mathbf{x}_i$. Those training samples with non-zero coefficient α_i are called support vectors, for the reason that they are essential to the solution of hypersphere $B(\mathbf{a}, R)$. The radius R is determined from the Karush-Kuhn-Tucker (KKT) conditions and can be obtained by calculating the distance from the center $\mathbf{a}$ to any support vector with $0 < \alpha_i < C$. The training samples with coefficient hitting the upper bound ($\alpha_i = C$) are considered as outliers. Finally, the class decision function $\phi(\mathbf{x})$ is equivalent to a function of the separating hypersphere and can be written as $\phi(\mathbf{x}) = \text{sign}(R^2 - ||\mathbf{x} - \mathbf{a}||^2)$.

2.2. *Support Vector Machine*

Suppose that we are given n training samples $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in R^d$ and $y_i \in \{+1, -1\}$. For example in Fig. 2(b), training samples are denoted as circle or square for two different classes. Based on the training set, a hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ could be obtained to separate the samples of different classes on the two sides of hyperplane, where $\mathbf{w}$ is the norm vector and b is the bias of the hyperplane. When training samples are linearly separable, support vector machine (SVM) yields the optimal hyperplane that separates two classes without training error, where optimization is in the sense of maximizing the minimum distance $(1/||\mathbf{w}||)$ from training samples to the hyperplane [Gunn, 1998][Burges, 1998][Chang & Lin, 2001]. In other words, the parameter pair $(\mathbf{w}, b)$ corresponding to the optimal hyperplane is a solution to the following optimization problem,

$$\begin{aligned} &\min : \frac{1}{2}||\mathbf{w}||^2, \\ &\text{subject to} : y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \ge 1. \end{aligned} \tag{3}$$

For linearly non-separable cases, the concept of a separating hyperplane is generalized by employing the slack variable ξ_i with potential training errors. Mathematically it can be written as

$$\begin{aligned} &\min : \frac{1}{2}||\mathbf{w}||^2 + C\sum_i \xi_i, \\ &\text{subject to} : \begin{cases} y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \ge 1 - \xi_i, \\ \xi_i \ge 0, \end{cases} \end{aligned} \tag{4}$$

where C is a parameter to adjust the weight of training errors in the minimization.

By introducing Lagrangian multipliers α_i, the above primal problem (4) is transformed to its dual form,

$$\begin{aligned} &\min : \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j - \sum_i \alpha_i, \\ &\text{subject to} : \begin{cases} \sum_i y_i \alpha_i = 0, \\ 0 \le \alpha_i \le C. \end{cases} \end{aligned} \tag{5}$$

The dual problem (5) can be solved via quadratic programming. After solutions of α_i are obtained from (5), the norm vector $\mathbf{w}$ is calculated as $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$, and the bias of the

Table 1. Comparison between SVDD and SVM.

	Support Vector Domain Description	**Support Vector Machine**
Objective	Description	Discrimination
Boundary	Hypersphere $B(\mathbf{a}, R)$	Hyperplane $(\mathbf{w}, b)$
Training samples	$\mathbf{x}_1, \ldots, \mathbf{x}_m$ where $\mathbf{x}_i \in R^d$	$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ $y_i \in \{+1, -1\}$
Primal problem	$\min : R^2$, subject : $\|\|\mathbf{x}_i - \mathbf{a}\|\|^2 \leq R^2$.	$\min : \frac{1}{2}\|\|\mathbf{w}\|\|^2$, subject : $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$.
Dual problem	$\sum_{i,j} \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j - \sum_i \alpha_i \mathbf{x}_i^T \mathbf{x}_i$, subject to : $\begin{cases} \sum_i \alpha_i = 1, \\ 0 \leq \alpha_i. \end{cases}$	$\frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j - \sum_i \alpha_i$, subject to : $\begin{cases} \sum_i y_i \alpha_i = 0, \\ 0 \leq \alpha_i. \end{cases}$
Solution	$\mathbf{a} = \sum_i \alpha_i \mathbf{x}_i$ $R^2 = \|\|\mathbf{x}_i - \mathbf{a}\|\|^2$ $i \in$ {Bounded Support Vector}	$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$ $b = y_i - \mathbf{w} \cdot \mathbf{x}_i$ $i \in$ {Bounded Support Vector}
Decision	$\text{sign}(R^2 - \|\|\mathbf{x} - \mathbf{a}\|\|^2)$	$\text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$

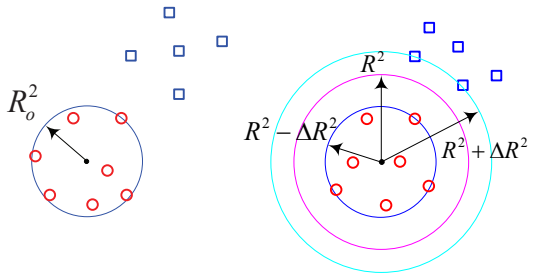
hyperplane b is determined from the KKT conditions. The class decision function can be expressed as $\phi(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$.

3. Margin Domain Integrated Classification (MDIC)

In this section, we will first introduce the concept of margin-domain integrated approach by using simple and illustrative examples for the purpose of revealing the principle. Then more general cases will be followed by the theory and applications.

3.1. *The Concept of MDIC*

We now introduce the margin and domain integrated concept. Before generalizing to multi-category cases, we change the class label from $\{+1, -1\}$ to $\{1, 2\}$ and reformulate the binary case to extend its capability. First let us consider the case of one-class with outliers as shown in Fig. 3(a). Given n training samples $(x_i, y_i)_{i=1}^n$ with $x_i \in R^d$ and $y_i \in \{1, 2\}$, the target class and outlier class are defined as $y_i = 1$ and $y_i = 2$, respectively. For simplicity the training samples are assumed to be spherically separable by a hypersphere $B(\mathbf{a}, R)$. If only the samples of the target class are used, one can find an optimal domain to describe the class through SVDD, which generates a hypersphere $B(a, R_{\text{O}})$ with center a and radius $R_{\text{O}} = \sqrt{R^2 - \Delta R^2}$. Now consider that outliers are involved in training the classifier as shown in Fig. 3(b). The goal is to generate a new hypersphere $B(a, R_{\text{X}})$ with a greater radius $R_{\text{X}} = \sqrt{R^2 + \Delta R^2}$ that pushes away the outliers. Between the two radiuses is the margin defined as ΔR^2. Heuristically, this idea can be explained as follows. When no outliers are involved, the hypersphere has to be as small as possible to avoid any misclassification of outliers as a target. When outliers become available in training, a more optimal hypersphere can be generated which should be between R_{O} and R_{X}.



(a) (b)

Fig. 3. (a) A domain for optimal description of the target class when no outliers are used for training, (b) the integration of domain and margin when both target and non-target class samples are used in training.

Our goal is to have the smallest hypersphere describing the target class and the largest hypersphere discriminating the outliers so that the margin is maximum. The boundary ultimately used for classification is between R_{O} and R_{X}, defined as $R = \sqrt{(R_{\mathrm{X}}^2 + R_{\mathrm{O}}^2)/2}$ in Fig. 3(b) where discrimination is incorporated into description. Let $R_{\mathrm{X}}^2 - R_{\mathrm{O}}^2 = 2\Delta R^2$. To minimize the cost of description, the radius R should be minimized. On the contrary, to minimize the cost of discrimination, the margin ΔR^2 should be maximized. The integration of the two capabilities can be achieved by minimizing the objective function $(\gamma R^2 - \Delta R^2)$, where the emphasis between the margin and description can be regulated by adjusting the importance between description or discrimination. When γ becomes larger, the purpose of optimization is shifted to description, closer to SVDD. When smaller, it is shifted to discrimination, closer to SVM. In addition, the above optimization is under the constraints of

$$\begin{cases} R^2 - \Delta R^2 \geq ||\mathbf{x}_i - \mathbf{a}||^2, \text{ for } i \in \mathcal{S}_1, \\ ||\mathbf{x}_i - \mathbf{a}||^2 \geq R^2 + \Delta R^2, \text{ for } i \in \mathcal{S}_2, \end{cases} \quad (6)$$

where $\mathcal{S}_j = \{i : y_i = j\}$. Constraints (6) simply imply that the samples of $\mathcal{S}_1$ are enclosed in the inner hypersphere and those of $\mathcal{S}_2$ are excluded outside of the outer hypersphere. The decision function can be written as $\phi(\mathbf{x}) = I(R^2 - ||\mathbf{x} - \mathbf{a}||^2 \geq 0) + 1$, where $I(\cdot)$ is the indicator function.

Let $f(\mathbf{x}) = R^2 - ||\mathbf{x} - \mathbf{a}||^2$, interpreted in Fig. 4(a) as contour lines which have the largest value R^2 at the center, value ΔR^2 at the inner hypersphere, value 0 at the decision hypersphere, and value $-\Delta R^2$ at the outer hypersphere. Constraints (6) can be written as

$$\begin{cases} f(\mathbf{x}_i) \geq \Delta R^2, \text{ for } i \in \mathcal{S}_1, \\ f(\mathbf{x}_i) \leq -\Delta R^2, \text{ for } i \in \mathcal{S}_2. \end{cases} \quad (7)$$

Based on the descriptive and discriminative nature of the binary MDIC, we extend its integration concept from binary to multiple classes. Mathematically, we want to find decision boundaries in vector space R^d based on the information from n training samples $\mathbf{x}_i$ to separate k different classes including $k-1$ target classes and one outlier class. We present multi-category MDIC in the forms of multiple-optimization and single-optimization, where multi-optimization MDIC constructs several binary classifiers and single-optimization MDIC solves the larger optimization problem in one step.

3.2. *Multi-Optimization MDIC*

Given $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, where $\mathbf{x}_i \in R^d$ is a training sample and $y_i \in \{1, 2, \ldots, k\}$ is the corresponding class label. For a k-category classification problem which contains $k-1$ target classes and one outlier class, we can find $k-1$ separating hyperspheres $B(\mathbf{a}_j, R_j)$ where all the samples of class j are enclosed by the inner hypersphere $B(\mathbf{a}_j, R_{\mathrm{O},j})$ and the other samples are excluded by the outer hypersphere $B(\mathbf{a}_j, R_{\mathrm{X},j})$. The radius $R_j = \sqrt{(R_{\mathrm{X},j}^2 + R_{\mathrm{O},j}^2)/2}$ and the margin $\Delta R_j = \sqrt{(R_{\mathrm{X},j}^2 - R_{\mathrm{O},j}^2)/2}$, for $j = 1, \ldots, k-1$. To reduce the multiclass problem to a set of binary problems, perhaps the simplest approach is to create one binary problem for each of the $(k-1)$ target classes. That is, for $j \in \{1, 2, \ldots, k-1\}$, we apply the given learning algorithm to a binary problem in which all examples labeled $y_i = j$ are considered positive examples and all other examples are considered negative examples. We then end up with $(k-1)$ hypotheses that somehow must be combined. We call this the one-versus-all (OVA) approach. Let $f_j(\mathbf{x}) = R_j^2 - ||\mathbf{x} - \mathbf{a}_j||^2$, for $j = 1, \ldots, k-1$. In the OVA fashion, the $k-1$ optimizations

(a) (b)

Fig. 4. Margin and domain integrated classification: (a) binary case, (b) multi-category case.

could be formulated as

$$\min : \gamma_j R_j^2 - \Delta R_j^2 + C\sum_{i=1}^{n}\xi_i,$$
$$\text{subject} : \begin{cases} f_j(\mathbf{x}_i) \geq \Delta R_j^2 - \xi_i, & \text{for } i \in \mathcal{S}_j, \\ f_j(\mathbf{x}_i) \leq -\Delta R_j^2 + \xi_i, & \text{for } i \notin \mathcal{S}_j, \\ \Delta R_j^2 \geq 0, \text{ and } \xi_i \geq 0, & \text{for } i = 1, \ldots, n. \end{cases} \quad (8)$$

Slack variables $\xi_i \geq 0$ are introduced and additional constraints $\Delta R_j^2 \geq 0$ are added in (8) to force the enclosing ball to be inside the excluding ball, which is not assured in the non-separable case.

After Lagrange transformation, (8) can be written and solved in the dual form of Lagrange multipliers α_{ij}. The center of hyperspheres $\mathbf{a}_j = \frac{1}{\gamma_j}\sum_{i=1}^{n}\alpha_{ij}y_i\mathbf{x}_i$, and the radius R_j can be determined from the KKT conditions. If we set $f_k(\mathbf{x}) = 0$, then the decision function $\phi(\mathbf{x}) = \arg\max_{\{1,\ldots,k\}} f_j(\mathbf{x})$.

3.3. *Single-Optimization MDIC*

Alternatively, to extend MDIC from binary to multiple-category in the single optimization fashion, constraints (6) can be written as

$$\begin{cases} f_1(\mathbf{x}_i) - f_2(\mathbf{x}_i) \geq \Delta R^2, \text{ for } i \in \mathcal{S}_1, \\ f_2(\mathbf{x}_i) - f_1(\mathbf{x}_i) \geq \Delta R^2, \text{ for } i \in \mathcal{S}_2, \end{cases} \quad (9)$$

where $f_2(\mathbf{x}) = 0$. The decision function for the binary case is expressed as $\phi(\mathbf{x}) = \arg\max(f_1, f_2)$.

For a k-category classification problem, $(k-1)$ separating hyperspheres $B(\mathbf{a}_j, R_j)$ could be found and the single optimization problem can be written as

$$\min : \sum_{u=1}^{k-1}(\gamma_u R_u^2 - \Delta R_u^2) + C\sum_{i=1}^{n}\xi_i$$
$$\text{subject} : \begin{cases} f_{y_i}(\mathbf{x}_i) - f_j(\mathbf{x}_i) \geq \Delta R_{y_i}^2 + \Delta R_j^2 - \xi_i, \\ \text{for } i = 1, \ldots, n, \\ \text{for } j = \{1, \ldots, k\} \setminus y_i, \\ \xi_i \geq 0, \text{ for } i = 1, \ldots, n, \\ \Delta R_j^2 \geq 0, \text{ for } j = 1, \ldots, k-1, \end{cases} \quad (10)$$

where $f_j(\mathbf{x}_i) = R_j^2 - ||\mathbf{x}_i - \mathbf{a}_j||^2$, for $j = 1, \ldots, k-1$, and $f_j(\mathbf{x}_i) = 0$, for $j = k$. The primal problem (10) is transformed to its dual form by employing Lagrange multipliers α_{ij}, and then the dual problem is solved by quadratic programming. After some calculations, we derive the dual problem in the following form.

$$\min_{\beta} : \beta^T\mathbf{Q}\beta + \mathbf{p}\beta$$
$$\text{subject to} : \begin{cases} \mathbf{1}_n^T\mathbf{F}_j\beta - \varphi_j = 1, \\ \mathbf{1}_n^T\mathbf{G}_j\beta = \gamma_j, \\ \mathbf{A}\beta \leq C\mathbf{1}_n, \end{cases} \quad (11)$$
$$\text{where} \begin{cases} \mathbf{Q} = \sum_{u=1}^{k-1}\frac{1}{\gamma_u}\mathbf{G}_u^T\mathbf{X}\mathbf{X}^T\mathbf{G}_u, \\ \mathbf{p} = \mathbf{1}_n^T\text{diag}(\mathbf{X}\mathbf{X}^T)\mathbf{G}_k, \\ \mathbf{A} = \mathbf{U} - \sum_{u=1}^{k}\mathbf{V}_u\mathbf{U}_u, \\ \mathbf{F}_j = \mathbf{V}_j\mathbf{U} + \mathbf{U}_j - 2\mathbf{V}_j\mathbf{U}_j, \\ \mathbf{G}_j = \mathbf{V}_j\mathbf{U} - \mathbf{U}_j, \end{cases}$$

for $j = 1, \ldots, k-1$. In (11), β is a vector containing the nk Lagrange multipliers. $\mathbf{V}_j$ is an $n \times n$ diagonal matrix with its diagonal element being 1 or 0 depending on $i \in \mathcal{S}_j$ or not. $\mathbf{U}_j$ is an $n \times nk$ matrix defined by $\mathbf{U}_j = (\mathbf{1}_k^j)^T \otimes \mathbf{I}_n$, where $\otimes$ is the Kronecker product and $\mathbf{1}_k^j$ is a k-dimensional vector with a 1 in its j-th element and 0 otherwise. $\mathbf{U} = \sum_{j=1}^{k}\mathbf{U}_j = \mathbf{1}_k^T \otimes \mathbf{I}_n$, where

$\mathbf{1}_k$ is a k-dimensional vector of 1 and $\mathbf{I}_n$ is the $n \times n$ identity matrix. After the solution of (11) is obtained, the center of the hypersphere is determined by $\mathbf{a}_j = \mathbf{X}^T\mathbf{G}_j\beta/\gamma_j$ and the radius R_j is determined by the KKT complementary conditions.

4. Experimental Results

To illustrate the concept and application of the MDIC approach, we first apply it to synthetic data in the first subsection. To evaluate the effectiveness, we then make comparisons between multi-category SVDD, SVM, and MDIC on three data sets from UCI machine learning repository in the second subsection. Finally, classification results on three standard MPEG-4 test video sequences are shown in the third subsection.

4.1. *Synthetic Data*

We conducted experiments on synthetic data of the two-dimensional space shown in Fig. 5. The training samples are distributed in a $[-1, 1] \times [-1, 1]$ square in which there are two target classes and one non-target class. Using the formulation of MDIC, we obtain the domain description shown in Fig. 5. The first row of Fig. 5 shows the results of using the linear kernel function $K(\mathbf{s}, \mathbf{t}) = \mathbf{s}^T\mathbf{t}$ and the second row shows the results of using the Gaussian kernel function $K(\mathbf{s}, \mathbf{t}) = \exp(-\frac{1}{2\sigma^2}||\mathbf{s} - \mathbf{t}||^2)$, where $\sigma = 0.5$. As can be seen, Gaussian kernel function fits the shape or distribution of training samples better than linear kernel function.

In the first column of Fig. 5, where γ_1 and γ_2 are both 1, the purpose of optimization is on the tight description for target classes, closer to SVDD. Therefore, the separating hypersphere coincides with the inner hypersphere and also with the outer hypersphere. In other words, no margin is allowed and miss rates would be large when generalizing to test samples. On the other hand, when γ_1 and γ_2 are close to 0, the purpose of optimization is shifted to discrimination, closer to SVM. In that case, false alarm rates would be large when generalization. The middle two columns of Fig. 5 illustrate how the two capabilities are alternated as the parameter varies.

4.2. *UCI Repository*

We investigated the performance of multi-category classifications on three data sets from UCI machine learning repository: vowel, Semeion handwritten digit, and letter recognition data sets [Asuncion & Newman, 2007]. Use of UCI machine learning repository is a common practice for comparison of different methods in the field of classification, which we follow such that a fair comparison between methods can be made. The statistics of data sets are listed in Table 2.

Table 2. Three data sets from the UCI repository.

Data set	# attributes	# classes	# samples
vowel	10	11	528
semeion	256	10	1593
letter	16	26	20000

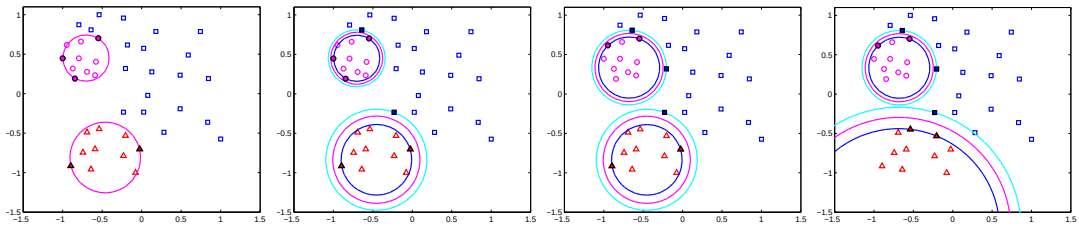
(i) **Vowel Recognition**
This data set contains 528 samples from speaker independent recognition of the 11 steady state vowels of British English. For each utterance, 10 floating-point numbers are provided as the attributes.

(ii) **Semeion Digit Recognition**
This data set consists of 1593 samples and 256 attributes. Each sample represents a handwritten digit, originally scanned with a resolution of one-byte (2^8) gray scale. After the scan, each pixel was transformed to 0 or 1 by threshold 127. Finally, each binary image was scaled into a 16x16 square box (the final 256 binary attributes).

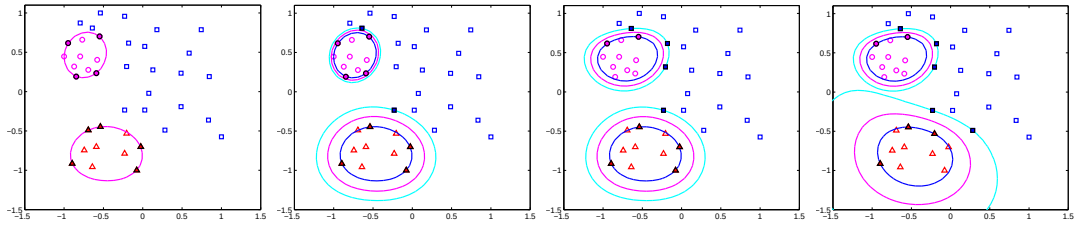
(iii) **Letter Image Recognition**
This data set contains 20000 samples, each corresponds to one of the 26 capital letters in the English alphabet. 16 integer-valued features such as statistical moments and edge counts are provided to represent each letter.

4.2.1. *Classification Accuracy*

Experiments were conducted on three cases of different target-to-outlier sample ratio: $\eta = 2/1, 1/1,$ and $1/2$. In the first case, we randomly pick 33 training samples from class one, another



(a)$\gamma_1 = 1.0, \gamma_2 = 1.0$ (b)$\gamma_1 = 0.9, \gamma_2 = 0.9$ (c)$\gamma_1 = 0.1, \gamma_2 = 0.9$ (d)$\gamma_1 = 0.1, \gamma_2 = 0.1$



(e)$\gamma_1 = 1.0, \gamma_2 = 1.0$ (f)$\gamma_1 = 0.9, \gamma_2 = 0.9$ (g)$\gamma_1 = 0.1, \gamma_2 = 0.9$ (h)$\gamma_1 = 0.1, \gamma_2 = 0.1$

Fig. 5. The effect of parameters γ on the resulted hyperspheres of MDIC. As γ increases, the classifier is closer to pure description which becomes discrimination as γ decreases. (a)-(d) linear kernel, (e)-(h) Gaussian kernel.

Table 3. Comparison of SVDD, SVM, and MDIC with regard to the total error rate, miss rate, and false alarm rate on data randomly extracted from UCI machine learning repository.

(i)	Total Error Rate P_T (%)			Miss Rate P_m (%)			False Alarm Rate P_f (%)		
Vowel	SVDD	SVM	MDIC	SVDD	SVM	MDIC	SVDD	SVM	MDIC
$\eta = 2/1$	22.5	5.5	**5.0**	18.2	**1.5**	5.3	20.6	7.4	**4.4**
$\eta = 1/1$	25.0	7.5	**3.0**	29.0	6.0	**5.0**	19.0	**1.0**	**1.0**
$\eta = 1/2$	18.5	5.5	**5.0**	30.6	**6.9**	**6.9**	11.7	2.3	**1.6**

(ii)	Total Error Rate P_T (%)			Miss Rate P_m (%)			False Alarm Rate P_f (%)		
Semeion	SVDD	SVM	MDIC	SVDD	SVM	MDIC	SVDD	SVM	MDIC
$\eta = 2/1$	13.0	4.0	**3.0**	19.7	**1.5**	**1.5**	**0.0**	7.4	4.4
$\eta = 1/1$	17.0	7.0	**6.0**	33.0	**7.0**	**7.0**	**1.0**	7.0	5.0
$\eta = 1/2$	14.5	5.5	**5.0**	40.3	**11.1**	**11.1**	**0.0**	2.3	1.6

(iii)	Total Error Rate P_T (%)			Miss Rate P_m (%)			False Alarm Rate P_f (%)		
Letter	SVDD	SVM	MDIC	SVDD	SVM	MDIC	SVDD	SVM	MDIC
$\eta = 2/1$	13.5	5.5	**4.5**	15.9	**0.0**	1.5	8.8	13.2	**7.4**
$\eta = 1/1$	18.5	8.0	**6.0**	33.0	**4.0**	5.0	**4.0**	9.0	6.0
$\eta = 1/2$	13.5	5.0	**3.5**	37.5	**8.3**	**8.3**	**0.0**	3.1	0.8

33 samples from class two, and the other 34 samples from the other classes. So the target-to-outlier sample ratio is nearly 2/1. In the second case, we randomly collect 25 samples from class one, 25 from class two, and 50 from the outlier class. For the 1/2 case, 18 samples from class one, 18 from class two, and 64 from outliers are contained in the training set.

In Table 3, we compare SVDD, SVM, and MDIC on the total error rate, miss rate and false alarm rate. The miss rate P_m is the percentage of target samples misclassified in the outlier class. The false alarm rate P_f is the rate of outlier samples misclassified as any one of the target classes. And the total error rate P_T is the percentage of any sample misclassified to other class.

The performance of MDIC is much better than SVDD since there is separating margin between boundaries of the target class and outliers in MDIC. The performance gap between the two is extremely large in miss rates. Since SVDD seeks the minimal domain of target classes for the tightest description, a target sample is easily misclassified as outliers when it falls outside of the boundary. Therefore, generalization performance of SVDD is not optimal especially when training samples do not contain all the variation of target samples. On the other hand, SVM has larger false alarm rates than MDIC. Since not all the variation of outlier samples could be possibly captured in the training, an outlier sample is easily misclassified as targets when it falls inside of the domain. Therefore, classification results of MDIC are better than those of SVM in the false alarm rates. Overall, MDIC has lower total error rates than SVDD and SVM.

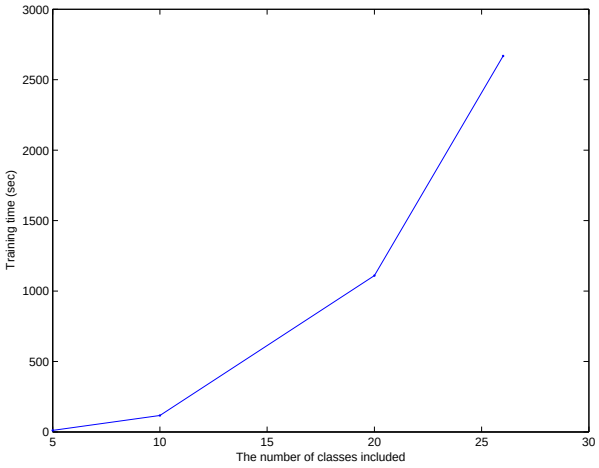


Fig. 6. Training time relates to the number of classes for solving multi-class problems in single optimization.

4.2.2. *Time of Computation*

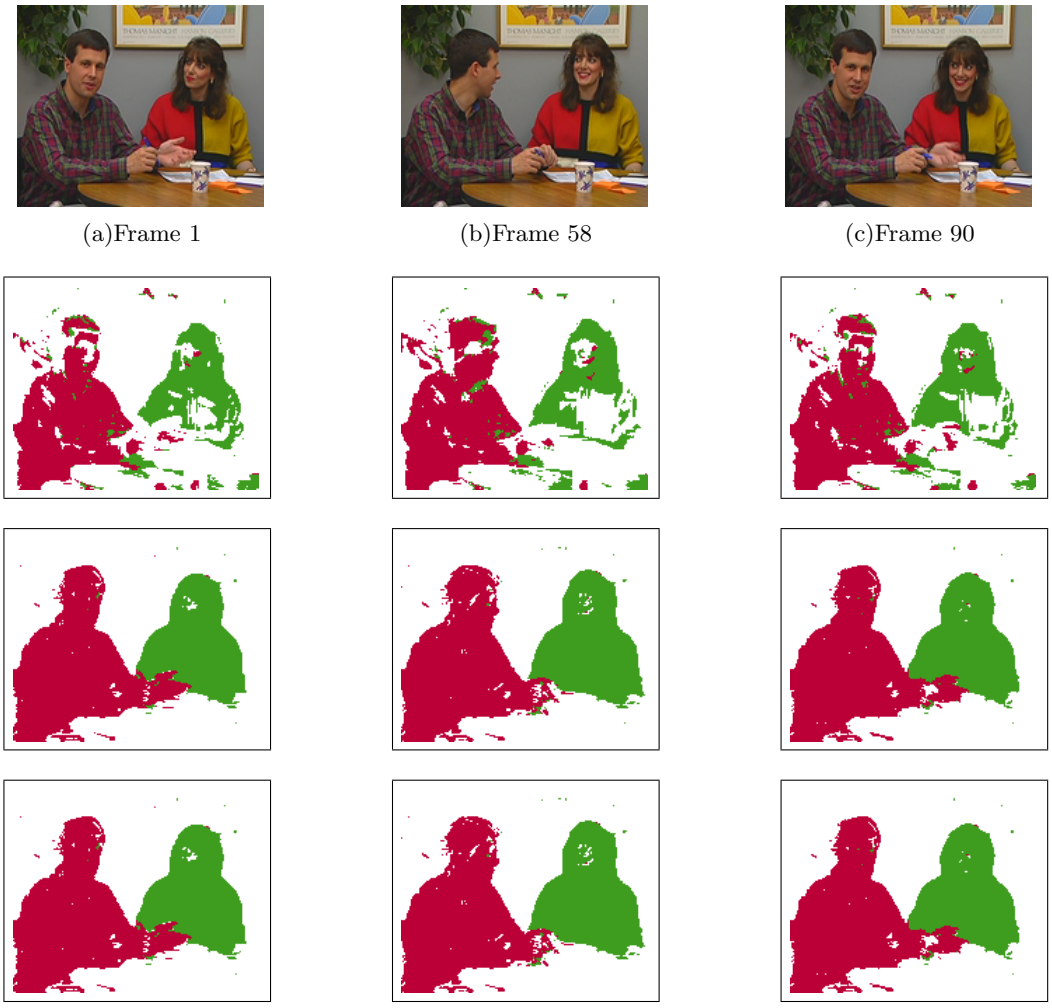
The computational cost of MDIC depends on the number of classes, and is in the same order as multi-SVM, for both training and classification. The formulation to solve multi-class problems in one step has variables proportional to the number of classes. Therefore, for multi-class methods, either several binary classifiers have to be constructed or a larger optimization problem is needed. Hence in general it is computationally more expensive to solve a multi-class problem than a binary problem with the same number of data.

To illustrate how the computation is related to the number of classes for solving multi-class problems in single optimization, 160 samples were randomly picked from the UCI letter image repository for training. The computer configuration for simulation in the laboratory is AMD Phenom(tm) Q9600B 2.30GHz Linux PC with 4GB RAM; the software adopted is IMSL C Library. As can be seen in Fig. 6, the training time increases rapidly as the number of classes increases. Using the polyfit function in MATLAB, the training-time curve could be described by $0.23k^3 - 2.82k^2 + 23.37k - 64.01$. In this case, the computation has $O(n^3)$ time complexity.

4.3. *Video Object Extraction*

Classification methods are used to extract video objects in this subsection. The purpose is to classify video objects such that video object based comparison can be made. Experiments were conducted on three standard MPEG-4 sequences: *Students*, *Trevor*, and *Sun Flower Garden* [Liu & Zheng, 2005][Liu & Zheng, 2008]. For each video sequence, the first frame was used in training, and another frame was used in validation and choosing the parameters $C \in \{0.1, 0.2, \ldots, 1.0\}$ and $\gamma \in \{0.05, 0.1, 0.2, \ldots, 1.0\}$. A third frame was used in testing which was totally unknown when training. 1/30 of the pixels in the training frame were randomly selected as the training samples and the block intensity information was used as the feature to represent each centering pixel.

In Figs. 7-9, performance comparisons were made between SVDD (row 1), SVM (row 2), and MDIC (row 3) in the order of three columns: column 1 for the training frame, column 2 for the validation frame, and column 3 for the test frame. On each row, the extracted objects could be differentiated by colors. The computational cost of MDIC depends on the size of the image for both training and classification.



(a)Frame 1 (b)Frame 58 (c)Frame 90

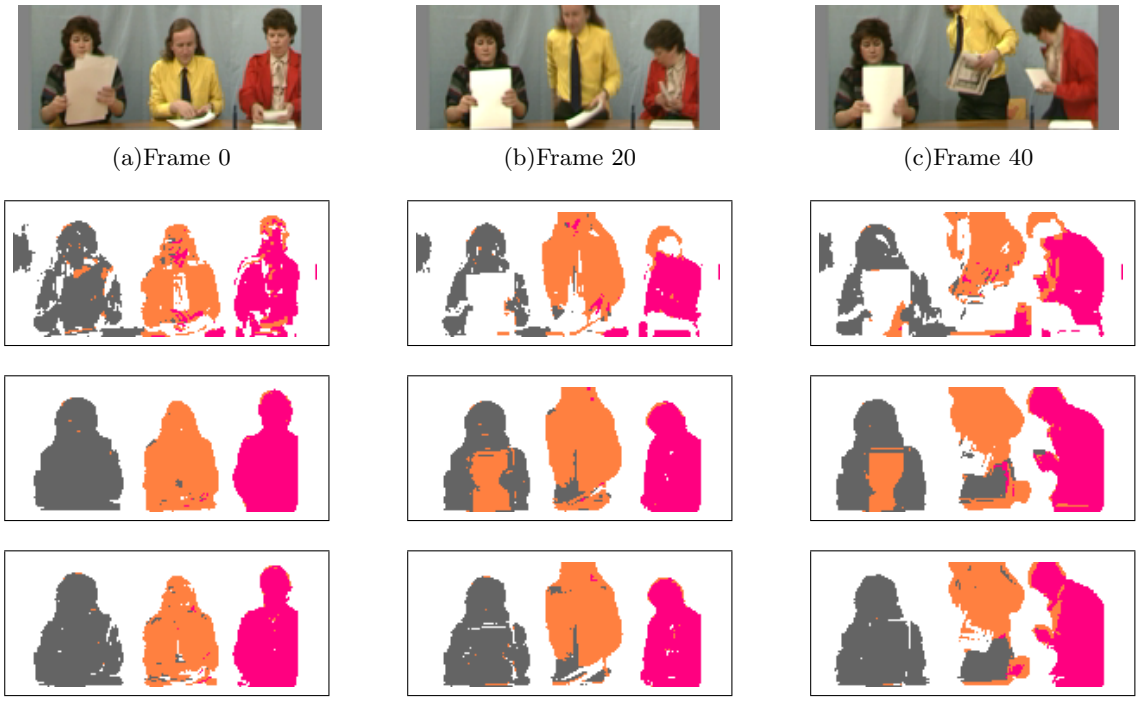
Fig. 7. The extraction performance of *Students*. First row: SVDD, second row: SVM, and third row: MDIC.

4.3.1. *Students*

As the major content of this sequence, the two students are chosen as two objects of interest, and along with the background this is a three-class classification problem. *Students* is a typical sequence of slow but heterogeneous motion. For example the male student turns the head and moves his hands while his body stays still most of the time. One can see that the proposed method works well, which discriminates the body parts of the students as well as their faces. The latter is not an easy task since the skin color is very similar between the two students. As can be seen, SVDD generates more misclassified pixels than those of MDIC and SVM.

4.3.2. *Trevor*

The three people in this video sequence are considered as three objects which makes it a four-class classification problem with the background as the fourth class. The original frames and the extracted objects are shown in Fig. 8. Unlike the *Students* sequence, the objects in this sequence change the appearance by a great deal. Taking the lady who sits at the farthest right as an example, her face changes from frontal to left-side view. Besides, the man in the middle is originally seated but finally standing. In Fig. 8, the performance of MDIC is better than those of SVDD and SVM as one can see that there are less classification errors on the third row.



(a)Frame 0 (b)Frame 20 (c)Frame 40

Fig. 8. The extraction performance of *Trevor*. First row: SVDD, second row: SVM, and third row: MDIC.

4.3.3. *Sun Flower Garden*

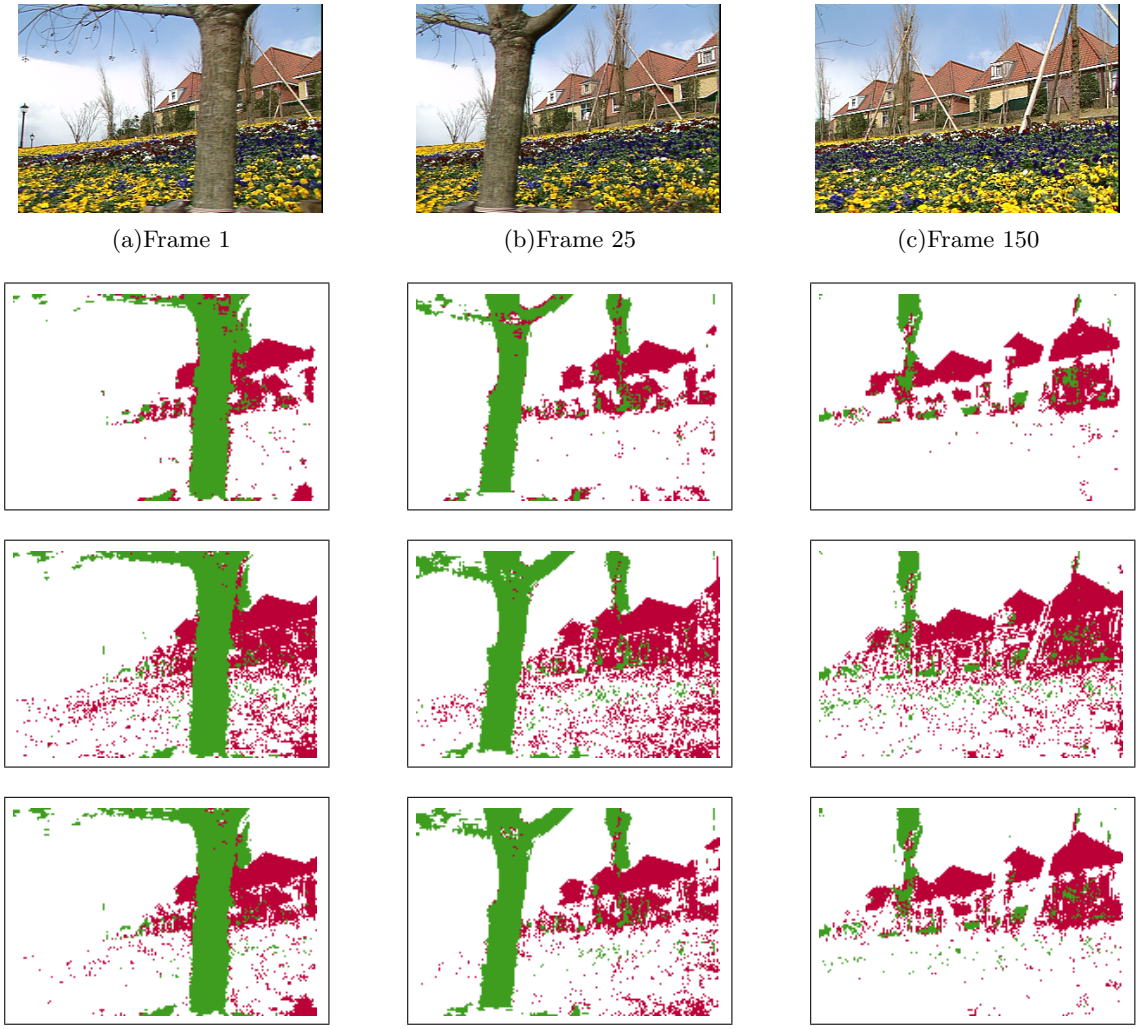
Different from the previous video conference kind of sequences, *Sun Flower Garden* displays a natural scene that is rich of colors and textures with a non-stationary camera. Each pixel is classified into one of the three categories: house, tree, and background. For the first few frames, the house is occluded by the tree. With the camera moving, the tree shifts toward the left side of the frame and finally disappears. As can be seen in Fig. 9, several pixels of the house were misclassified as background by SVDD on the first row. It is especially noticeable in Frame 150 when the house spreads out in the whole frame due to the motion of the camera. On the other hand, many background pixels were misclassified as objects by SVM on the second row, which implies higher false alarm rates than those of the previous rows. Overall, MDIC has lower error rates on the test frame.

5. Conclusion

This paper presents a multi-category classification approach which is based on the classical SVM and SVDD concepts but represents a significant departure from the two methods. Although SVM is optimized for a large margin between different classes, it does not evaluate the fitness of a sample to a particular class. Therefore, the problem results in potentially high false-alarm rates. On the other hand, SVDD is optimized for describing a minimum domain based on the training samples. However, the domain is tight and no separation occurs. As a result, it causes elevated miss rates in many emerging application. Inspired by the binary learning machine MEMEM, a multi-category classifier which integrates description and discrimination is proposed in this paper to overcome the deficiency of SVM and SVDD but to retain the advantages associated with each method.

In this paper, a novel multi-category classifier called margin and domain integrated classifier (MDIC) is developed for its capability of both discrimination and description. The proposed approach has the following features:

- The notions of domain description and separation margin are integrated in a single model for enhancing classification accuracy in multiple classes.



(a)Frame 1 (b)Frame 25 (c)Frame 150

Fig. 9. The extraction performance of *Sun Flower Garden*. First row: SVDD, second row: SVM, and third row: MDIC.

- A domain measure γ is provided for each target class to adjust the regulation between the class description and the discrimination from other target class or outliers.
- MDIC treats multi-class classification jointly rather than sequentially to achieve an optimal performance globally.

The ratio between description and discrimination is flexible so that an optimal performance is achieved to adapt various purposes and conditions of each application. Experiments were conducted and promising results have been obtained. Future work will encompass research efforts on lowering computational complexity of MDIC when the number of class is huge.

References

Asuncion, A. and Newman, D. J. [2007] "UCI machine learning repository," [`http://archive.ics.uci.edu/ml/`]. School of Information and Computer Science, University of California, Irvine, CA.

Banerjee, A., Burlina, P. and Meth, R. [2007] "Fast hyperspectral anomaly detection via SVDD," *Proc. Intl. Conf. on Image Processing*, **4**, 101–104.

Burges, C. [1998] "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, **2**, 121–167.

Chang, C.-C. and Lin, C.-J. [2001] LIBSVM: a library for support vector machines, [`http://www.csie.ntu.edu.tw/~cjlin/libsvm`].

Cortes, C. and Vapnik, V. [1995] "Support-vector networks", *Machine Learning*, **20**(3), 273–297.

Duda, R. O., Hart, P. E. and Stork, D. G. [2001] *Pattern Classification*, 2nd Ed. (John Wiley & Sons).

Goh, K.-S., Chang, E. Y. and Li, B. [2005] "Using one-class and two-class SVMs for multiclass image annotation," *IEEE Trans. on Knowledge and Data Engineering*, **17**(10), 1333–1346.

Gunn, S. R. [1998] "Support vector machines for classification and regression," Technical Report, School of ECE, University of Southampton.

Hastie, T., Tibshirani, R. and Friedman, J. [2001] *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer).

Hsu, C.-W. and Lin, C.-J. [2002] "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, **13**(2), 415–425.

Joshi, A., Porikli, F. and Papanikolopoulos, N. [2010] "Breaking the interactive bottleneck in multiclass classification with active selection and binary feedback," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2995–3002.

Kim, T. and Jo, K. -H. [2009] "Real-time object detection using two background models under shaking camera," *Intl. J. Information Acquisition*, **6**(1), 13–21.

Lai, C., Tax, M. J., Duin, R., Zbieta, E., Ekalska, P. and Ik, P. [2004] "A study of combined image representation for image classification and retrieval," *Intl. J. Pattern Recognition and Artificial Intelligence*, **18**, 867–890.

Lee, Y. and Lin, Y. and Wahba, G. [2004] "Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data," *J. American Statistical Association*, **99**(465), 67–81.

Li, M. and Hao, Y. [2008] "Accelerated kernel CCA plus SVDD: a three-stage process for improving face recognition," *J. Computers*, **3**, 94–100.

Lin C.-F. and Wang, S.-D. [2002] "Fuzzy support vector machines," *IEEE Trans. Neural Networks*, **13**(2), 464–471.

Liu, Y., Shen, X. and Doss, H. [2005] "Multicategory ψ-learning and support vector machine: computational tools," *J. Computational and Graphical Statistics*, **14**, 219-236.

Liu, Y. and Zheng, Y. F. [2005] "One-against-all multi-class SVM classification using reliability measures," in *Proc. Intl. Joint Conf. Neural Network*, **2**, 849–854.

Liu, Y. and Zheng, Y. F. [2005] "Video object segmentation and tracking using ψ-learning classification," *IEEE Trans. Circuits Syst. Video Tech.*, **15**(7), 885–899.

Liu, Y. and Zheng, Y. F. [2006] "Minimum enclosing and maximum excluding machine for pattern description and discrimination," in *Proc. Intl. Conf. Pattern Recognition*, **3**, 129–132.

Liu, Y., Zheng, Y. F. and Shen, X. [2008] "Applying the multi-category learning to multiple video object extraction," *Pattern Recognition*, **41**, 2777–2788.

Osuna, E., Freund, R. and Girosi, F. [1997] "Support vector machines: training and applications," *AI Memo* 1602, Massachusetts Institute of Technology.

Platt, J. C. [1999] "Probabilistic outputs for SVMs and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers* (MIT Press).

Pontil, M. and Verri, A. [1998] "Support vector machines for 3D object recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **20**(6), 637–646.

Rifkin, R. and Klautau, A. [2004] "In defense of one-vs-all classification," *J. Machine Learning Research*, **5**, 101-141.

Scholkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. and Williamson, R. C. [2001] "Estimating the support of a high-dimensional distribution," *Neural Computation*, **13**(7), 1443–1471.

Tax, D. M. J. and Duin, R. P. W. [1999] "Support vector domain description," *Pattern Recognition Letters*, **20**(11-13), 1191–1199.

Tax, D. M. J. [2009] "DDtools: a Matlab toolbox for data description, outlier and novelty detection," [`http://homepage.tudelft.nl/n9d04/dd_tools.html`], version 1.7.3.

Vapnik, V. [1999] *The Nature of Statistical Learning Theory*, 2nd Ed. (Springer-Verlag, New York).

Weston, J. and Watkins, C. [1999] "Support vector machines for multi-class pattern recognition," in *Proc. 7th European Symposium on Artificial Neural Networks*, 219–224.

Yuan, C. and Casasent, D. [2003] "A novel support vector classifier with better rejection performance," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, **1**, 419–424.

Zeng, Z., Fu, Y., Roisman, G., Wen, Z., Hu, Y. and Huang, T. S. [2006] "One-class classification for spontaneous facial expression analysis," *Proc. 7th Intl. Conf. Automatic Face and Gesture Recognition*, 281–286.