Direct Aspect-Based 3-D Object Recognition

Massimiliano Pontil and Alessandro Verri

INFM - Dipartimento di Fisica dell'Università di Genova,
Via Dodecaneso 33, 16146 Genova (I)

Abstract. In this paper a method for 3-D object recognition based on *Support Vector Machines* (SVM) is proposed. Given a set of points which belong to either of two classes, a SVM finds the hyperplane that leaves the largest possible fraction of points of the same class on the same side, while maximizing the distance of the closest point. Recognition with SVMs does not require feature extraction and can be performed directly on images regarded as points of an N-dimensional object space. The potential of the proposed method is illustrated on a database of 7200 images of 100 different objects. The excellent recognition rates achieved in all the performed experiments indicate that the method is well-suited for aspect-based recognition.

1 Introduction

Recently, aspect-based recognition strategies have received increasing attention from both the psychophysical [9, 4] and computer vision [8, 2, 6] communities. This is mainly due to the fact that these strategies appear to be well-suited for the solution of recognition problems in which geometric models of the viewed objects can be difficult, if not impossible, to obtain.

In this paper an aspect-based method for the recognition of 3-D objects is proposed. The method is based on Support Vector Machines. Given a set of points which belong to either of two classes, a SVM looks for the hyperplane that leaves the largest possible fraction of points of the same class on the same side, while maximizing the distance from the closest point. According to [10], given fixed but unknown probability distributions, a SVM minimizes the risk of misclassifying not only the examples in the training set but also the *yet to be seen* examples of the test set. As opposed to other aspect-based methods, recognition with SVMs does not require feature extraction and can be performed directly on images regarded as points of an N-dimensional object space. The high dimensionality of the object space makes SVMs very effective decision surfaces, while the recognition stage is essentially reduced to deciding on which side of a hyperplane lies a given point in object space.

The aim of this paper is to illustrate the potential of recognition techniques that operate directly on grey level images. The proposed method has been tested on the COIL database[1] consisting of 7200 images of 100 objects. Half of the images were used as training examples, the remaining half as test images. We

[1] The images of the COIL database (Columbia Object Image Library) can be downloaded through anonymous ftp from www.cs.columbia.edu.

discarded color information and tested the method on registered images (as obtained directly from COIL) and on the same images corrupted by synthetically generated random noise. The remarkable recognition rates achieved in all the performed experiments indicate that the method is well-suited for aspect-based recognition. Preliminary comparisons with other recognition methods, like perceptrons, show that the proposed method is far more robust in the presence of noise.

The paper is organized as follows. In Section 2 we review the main concepts of the theory of SVMs in the linear case. Then, in Section 3, we report the experiments of 3-D object recognitin with a linear SVM. Finally, Section 4 summarizes the conclusions that can be drawn from the presented research.

2 Theoretical overview

In this Section we recall the basic definitions of the theory of linear SVMs [10, 3].

2.1 Optimal separating hyperplane

In what follows we assume we are given a set S of points $\mathbf{x}_i \in \mathbb{R}^n$ with $i = 1, 2, \ldots, N$. Each point $\mathbf{x}_i$ belongs to either of two classes and thus is given a label $y_i \in \{-1, 1\}$. The goal is to establish the equation of a hyperplane that divides S leaving all the points of the same class on the same side while maximizing the distance between the two classes. To this purpose we need some preliminary definitions.

Definition 1. The set S is *linearly separable* if there exist $\mathbf{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \tag{1}$$

for $i = 1, 2, \ldots, N$.

The pair $(\mathbf{w}, b)$ defines a hyperplane of equation

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

named *separating hyperplane* (see Figure 1(a)). If we denote with w the norm of $\mathbf{w}$, the signed distance d_i of a point $\mathbf{x}_i$ from the separating hyperplane $(\mathbf{w}, b)$ is given by

$$d_i = \frac{\mathbf{w} \cdot \mathbf{x}_i + b}{w}. \tag{2}$$

Combining inequality (1) and equation (2), for all $x_i \in S$ we have

$$y_i d_i \geq \frac{1}{w}. \tag{3}$$

Therefore, $1/w$ is the lower bound on the distance between the points $\mathbf{x}_i$ and the separating hyperplane $(\mathbf{w}, b)$.

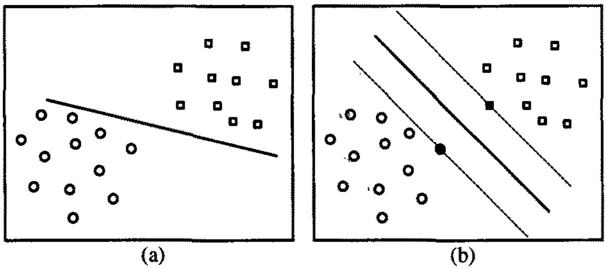

Fig. 1. Separating hyperplane (a) and OSH (b). The dashed lines in (b) identify the margin.

Definition 2. Given a separating hyperplane $(\mathbf{w}, b)$ for the linearly separable set S, the *canonical representation* of the separating hyperplane is obtained by rescaling the pair $(\mathbf{w}, b)$ into the pair $(\mathbf{w}', b')$ in such a way that the distance of the closest point, say $\mathbf{x}_j$, equals $1/w'$.

Through this definition we have that

$$\min_{\mathbf{x}_i \in S} \{y_i(\mathbf{w}' \cdot \mathbf{x}_i + b')\} = 1.$$

Consequently, for a separating hyperplane in the canonical representation, the bound in inequality (3) is tight. In what follows we will assume that a separating hyperplane is always given a canonical representation and thus write $(\mathbf{w}, b)$ instead of $(\mathbf{w}', b')$. We are now in a position to define the notion of OSH.

Definition 3. Given a linearly separable set S, the *optimal separating hyperplane* is the separating hyperplane which maximizes the distance of the closest point of S.

Since the distance of the closest point equals $1/w$, the OSH can be regarded as the solution of the problem of minimizing $1/w$ subject to the constraint (1), or

Problem **P1**
Minimize $\frac{1}{2}\mathbf{w} \cdot \mathbf{w}$
subject to $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, i = 1, 2, \ldots, N$

Note that the parameter b enters in the constraints but not in the function to be minimized. The quantity $2/w$, the lower bound of the minimum distance between points of different classes, is named *margin*. Hence, the OSH can also be seen as the separating hyperplane which maximizes the margin (see Figure 1(b)). We now study the properties of the solution of the Problem **P1**.

2.2 Support vectors

Problem **P1** is usually solved by means of the classical method of Lagrange multipliers. For more details and a thorough review of the method see [1]. Here we simply summarize the main results.

If we denote with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ the N nonnegative Lagrange multipliers associated with the constraints (1), we have

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i,$$

while b can be determined from $\boldsymbol{\alpha}$ and the Kühn-Tucker conditions

$$\alpha_i \left(y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \right) = 0, \quad i = 1, 2, \ldots, N. \tag{4}$$

Note that the only $\bar{\alpha}_i$ that can be nonzero in equation (4) are those for which the constraints (1) are satisfied with the equality sign. This has an important consequence. Since most of the $\bar{\alpha}_i$ are usually null, the vector $\bar{\mathbf{w}}$ is a linear combination of a relatively small percentage of the points $\mathbf{x}_i$. These points are termed *support vectors* because they are the closest points from the OSH and the only points of S needed to determine the OSH (see Figure 1(b)).

Given a support vector $\mathbf{x}_j$, the parameter $\bar{b}$ can be obtained from the corresponding Kühn-Tucker condition as

$$\bar{b} = y_j - \bar{\mathbf{w}} \cdot \mathbf{x}_j. \tag{5}$$

The problem of classifying a new data point $\mathbf{x}$ is now simply solved by looking at the sign of

$$\bar{\mathbf{w}} \cdot \mathbf{x} + \bar{b}.$$

Therefore, the support vectors condense all the information contained in the training set S which is needed to classify new data points.

2.3 Nonseparable case

If the set S is not linearly separable or one simply ignores whether or not the set S is linearly separable, the problem of searching for an OSH is meaningless (there may be no separating hyperplane to start with). Fortunately, the previous analysis can be generalized by introducing N nonnegative variables $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_N)$ such that

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, N. \tag{6}$$

The purpose of the variables ξ_i is to allow for a small number of misclassified points. If the point $\mathbf{x}_i$ satisfies inequality (1), then ξ_i is null and (6) reduces to (1). Instead, if the point $\mathbf{x}_i$ does not satisfy inequality (1), the extraterm $-\xi_i$ is added to the right hand side of (1) to obtain inequality (6). The generalized OSH is then regarded as the solution to

Problem **P2**
Maximize $\frac{1}{2}\mathbf{w} \cdot \mathbf{w} + C \sum \xi_i$
subject to $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \; i = 1, 2, \ldots, N$
$\boldsymbol{\xi} \geq 0.$

The purpose of the extraterm $\sum \xi_i$, where the sum is for $i = 1, 2, \ldots, N$, is to keep under control the number of misclassified points. Note that this term leads to a more robust solution, in the statistical sense, than the intuitively more appealing term $\sum \xi_i^2$. In other words, the term $\sum \xi_i$ makes the OSH less sensitive to the presence of outliers in the training set.

In analogy with what was done for the separable case, Problem **P2** can be solved by means of the method of Lagrange multipliers. If we denote with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ the N nonnegative Lagrange multipliers associated with the constraints (6), we find

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i,$$

while b can again be determined from $\boldsymbol{\alpha}$ and the new Kühn-Tucker conditions

$$\alpha_i \left(y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 + \xi_i \right) = 0 \tag{7}$$

$$(C - \alpha_i)\xi_i = 0 \tag{8}$$

Similarly to the separable case, the points $\mathbf{x}_i$ for which $\bar{\alpha}_i > 0$ are termed *support vectors.* The main difference is that here we have to distinguish between the support vectors for which $\bar{\alpha}_i < C$ and those for which $\bar{\alpha}_i = C$. In the first case, from condition (8) it follows that $\bar{\xi}_i = 0$, and hence, from condition (7), that the support vectors lie at a distance $1/\bar{w}$ from the OSH. In practice, these support vectors are the same of the separable case. The support vectors for which $\bar{\alpha}_i = C$, instead, are either misclassified points (if $\xi_i > 1$), or points correctly classified but closer than $1/\bar{w}$ from the OSH (if $0 < \xi \leq 1$).

From the computational viewpoint, the determination of the support vectors and associated OSH requires the solution of a problem of quadratic programming. In this research we have adopted an implementation of the *complementary pivoting algorithm* which is applied to the *linearly complementary problem* naturally associated with the dual of the quadratic programming problem **P2**. For more details see [1].

Concluding, we point out that the entire construction can also be extended rather naturally to include nonlinear separating surfaces [10]. However, since for the research described in this paper this extension was not needed, we do not further discuss this issue here.

3 Experiments

In this section we first discuss the recognition system and then illustrate its performance in different working conditions.

The recognition system accepts as input one of the 7200 images of the COIL database. This database contains 72 images (24 bits for each of the RGB channels and 128 × 128 pixels) of 100 objects positioned in the center of a turntable and observed from a fixed viewpoint. For each object, the turntable is rotated of 5° per image. Figures 2 and 3 show a selection of the objects in the database and one every three views (or images) of a specific object respectively. As explained in detail in [6], the object region is re-sampled so that the larger of the two dimensions fits the image size. Consequently, the apparent size of an object may change considerably from image to image, especially for the objects which are not symmetric with respect to the turntable axis.

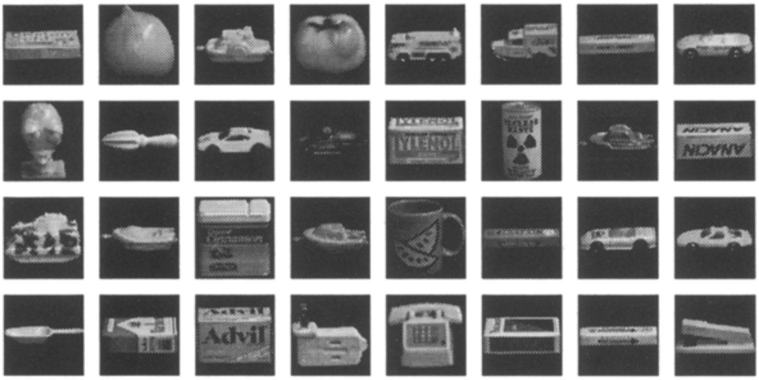

Fig. 2. Images of 32 of the COIL objects.

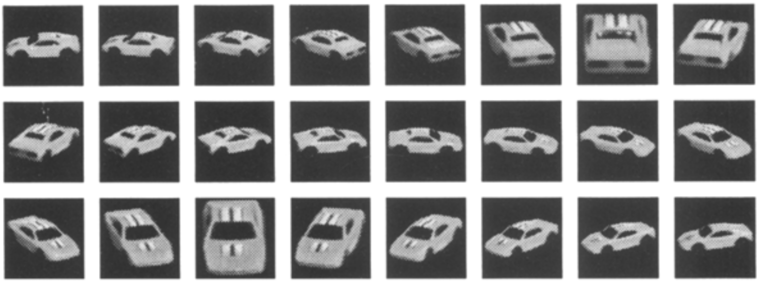

Fig. 3. Twentyfour of the 72 images of a COIL object.

Each image was transformed in a black and white 8 bit image through the usual conversion formula between RGB and grey level and rescaling the range between the minimum and maximum value. Finally, the image resolution was reduced to 32 × 32 by simply averaging the grey values over 4 × 4 pixel windows. The aim of these transformations was to reduce the dimensionality of the representation given the relatively small number of images available.

In the first series of experiments we considered groups of 32 objects and formed training and test set of equal size. Either set contained 36 images (one every 10°) of each of the 32 objects. Each image was regarded as a vector of

$32\times 32 = 1024$ components, and for each pair of objects i and j, $i, j = 1, 2, \ldots, 32$, the OSH was computed using only the $36 + 36 = 72$ images of the training set corresponding to objects i and j. Typically, we have found a number of support vectors ranging between 30% and 60% of the 72 training images for each object pair. This large percentage of support vectors (far above the percentage predicted by the theory) is easily explained by the high dimensionality of the object space combined with the small number of examples.

Recognition was performed following the rules of a tennis tournament. Given a previously unseen image and a database of N possible objects, each object is regarded as a *player*, and in each *match* the system temporarily classifies the previously unseen image according to the OSH relative to the pair of players involved in the match. The losing players are out and only in the final match the classification is unambiguously determined. This procedure requires $N - 1$ classifications (if N is the number of players). Since the classification stage is simple and fast, this does not seem to be crucial.

For many random choices of 32 of the 100 objects the system reached perfect score. Therefore, we decided to select by hand the 32 *more difficult* objects. By doing so the system finally mistook a packet of chewing gum for another very similar packet of chewing gum (in very similar pose) in one case.

In a second series of experiments we perturbed the original data by adding zero mean random noise to the grey value of each pixel and rescaling the grey levels between 0 and 255. The system performed equally well on noise corrupted images for maximum noise up to ± 75 grey levels and degrades gracefully for higher percentages of noise (see Table 1). It must be noted that most of the errors are usually due the three chewing gum packets of Figure 2 which become practically indistinguishable as the noise increases. Clearly, the very good statistics of Table 1 are partly due to the "filtering effects" of the averaging stage described in the previous section.

Table 1. Average overall recognition rates (gl = grey levels).

	no added noise	$\pm 25gl$	$\pm 50gl$	$\pm 100gl$	$\pm 200gl$
av.rec.rates	99.9%	99.8%	99.2%	98.4%	93.8%

In summary, the proposed method performs recognition with excellent percentages of success even in the presence of very similar objects. From the obtained experimental results, it can easily be inferred that the method achieves very good recognition rates because the OSH maximizes the margin and hence is able to produce remarkable classification performances even in the presence of large amount of noise. Preliminary comparisons with other recognition methods, like perceptrons, confirm the robustness of the proposed method. The average perceptron finds rather easily a separating hyperplane between the points of the training set (as can be expected from the fact that the problem is to separate a small number of points in a space of high dimension). However, the recognition rates of the perceptron fall down very quickly in the presence of even small perturbation of the data in the test set. It is worthwhile noticing that while the recognition time is practically negligible, the training stage (in which all the $32 \times 31/2 = 496$ OSHs must be determined) takes about 15 minutes on a SPARC10 workstation.

4 Conclusion

In this paper we have proposed a method for the recognition of 3-D objects from a single view based on support vector machines. As predicted by the theory of SVMs, it appears that the method can be effectively *trained* even if the number of examples is much lower than the dimensionality of the object space. This agrees with the theoretical expectation that can be derived by means of VC-dimension considerations [10]. The remarkably good results which we have described indicate that the method is likely to be very useful for direct 3-D object recognition.

Clearly, much work remains to be done. Currently we are looking at the stability of the method with respect to other form of perturbation, like brightness and light changes. Presumably the most difficult open problem is to what extent the method can be adjusted to tolerate occlusions.

References

1. Bazaraa, M. and Shetty, C.M. *Nonlinear programming* (John Wiley, New York, 1979).
2. Brunelli, R. and Poggio, T. 1993. "Face Recognition: Features versus Templates," *IEEE Trans. on PAMI* **15**: 1042-1052.
3. Cortes, C. and Vapnik, V. 1995. "Support Vector Network," *Machine learning* **20**: 1-25.
4. Edelman, S., Bulthoff, H., and Weinshall, D. 1989. "Stimulus Familiarity Determines Recognition Strategy for Novel 3-D Objects," AI Memo No. 1138, AI Lab, MIT.
5. Huttenlocher, D.P., Klanderman, G.A., and Rucklidge, W.J. 1993. "Comparing Images Using the Hausdorff Distance," *IEEE Trans. on PAMI* **15**:850-863.
6. Murase, H. and Nayar, S.K. 1995. "Visual Learning and Recognition of 3-D Object from Appearance," *Int. J. Comput. Vision* **14**:5-24.
7. Osuna, E., Freund, R.,and Girosi, F. "Training Support Vector Machines: an Applications to Face Detection." (Submitted to CVPR97).
8. Poggio, T. and Edelman, S. 1990. "A Network that Learns to Recognize Three-Dimensional Objects," *Nature*, Vol. 343, pp. 263-266.
9. Tarr, M. and Pinker, S. 1989. "Mental Rotation and Orientation-Dependence in Shape Recognition," *Cognitive Psychology*, Vol. 21, pp. 233-282.
10. Vapnik, V. *The Nature of Statistical Learning Theory* (Springer-Verlag, New York, 1995).