How Does the Visual System Represent Objects for Recognition?

Marvin Theiß
Student ID: [redacted]

S1: Visual Object Perception
MBB–MA–THM–3: Visual Cognition and Object Perception

February 24, 2023

Abstract

Shape-based object recognition theories propose that the visual system compares the two-dimensional image of a viewed object to representations stored in memory to recognise an object. Recognition is declared if a representation is found that closely matches the image projected onto the retina. While agreeing on this overall approach, recognition theories based on shape differ significantly regarding how representations are stored in memory and how these are matched to the images of viewed objects. For example, these theories can be differentiated by their degree of view-dependency and the dimensionality of the stored representations. Here, I review a study by Bülthoff and Edelman published in 1992, which suggests that the visual system likely represents objects by multiple 'two-dimensional snapshots' and uses an approximation scheme for recognition.

Keywords: object recognition, view alignment, linear combination of views, view approximation

Introduction

Every time we look at a scene, we recognise objects instantly and, apparently, almost effortlessly. Yet, from a computational perspective, the task of recognising objects is inherently difficult. Part of this is to do with the fact that the two-dimensional image projected onto the retina by a three-dimensional object can vary drastically depending on the object's orientation, its size, and the illumi-

nation of the scene. From this point of view, shape-based object recognition can be interpreted as the difficult task of inverting a one-to-many mapping.

Theories of object recognition based on shape generally suggest that this task is performed by comparing a viewed object to representations stored in memory. As pointed out by Tarr and Kriegman (2001), such an approach raises three questions: First, what do the representations stored in memory look like? Second, how are these representations generated in the first place? Third, how are viewed objects matched to these mental representations to recognise objects?

Various schemes of shape-based object recognition provide different answers to these questions. For example, some theories postulate the storing of multiple two-dimensional representations (*view-dependent*). In contrast, other theories suggest just a single three-dimensional representation (*view-independent*) being stored in memory. Similarly, different underlying mechanisms have been proposed for the 'matching task' of object recognition, such as decomposition into parts, structural descriptions, alignment schemes and approximation, to name a few.

Here, I will focus on three approaches to shape-based object recognition: recognition by alignment (Ullman, 1989), linear combination of views (Ullman & Basri, 1989), and view approximation (Poggio & Edelman, 1990). In the first part of this essay, I will introduce these theories one by one. In the second part, I will elaborate on a psychophysical experiment conducted by Bülthoff and Edelman (1992) that tested these three theories based on their predicted capability to generalise from learned to unfamiliar views.

Shape-Based Object Recognition Theories

The experiment conducted by Bülthoff and Edelman (1992) plots three theories of object recognition against each other: recognition by alignment, linear combination of views, and view approximation. These three theories predict different patterns of recognition performance when generalising from familiar to unfamiliar views, as I will explain next.

Recognition by Alignment. The first model of object recognition I will elaborate on is an *alignment* approach proposed by Ullman (1989), titled 'alignment of pictorial descriptions'. According to Ullman, the task of recognising a viewed object can be con-

sidered a search for the closest match in a vast space comprised of stored object models and their possible views. Recognition by alignment splits this search into two separate stages: First, a unique transformation is computed for every view of every object model stored in memory that optimally aligns the stored model and the viewed object. This part of the process is called the 'alignment' stage. Second, all aligned models are compared with the viewed object, and recognition is declared for the model that best matches the viewed object according to some 'measure of fit' (Ullman, 1989).

This process can formally be expressed as follows: Let $X^{(2D)}$ denote the two-dimensional image of the viewed object projected onto the retina. Further, let $(M_i)_{i\in I}$ and $(T_{ij})_{i\in I, j\in J}$ denote the sets of stored models and allowed transformations thereof, respectively. The alignment stage can then be formulated as the optimisation problem

$$T_i = \underset{(T_{ij})_{j\in J}}{\operatorname{argmin}} \left\| T_{ij}(M_i) - X^{(2D)} \right\|, \quad i \in I,$$

where $T_{ij}(M_i)$ denotes the application of the transformation T_{ij} to the model M_i and the norm $\|\bullet\|$ is used as a measure of dissimilarity. Note that underlying the computation of the best-fitting transformation T_i is the idea that "objects have certain invariant properties [or 'feature points'] common to all of their views" (Ullman, 1989, p. 201) and that these features can be relied on to align two views of the same object properly.

Once the unique transformation T_i has been established for every model M_i, recognition is declared for the model

$$M = \underset{(M_i)_{i\in I}}{\operatorname{argmin}} \left\| T_i(M_i) - X^{(2D)} \right\|$$

that best matches the two-dimensional image of the viewed object.

In his work, Ullman (1989) offers two alternative approaches the alignment scheme uses, which differ in their degree of view-dependency. One suggestion is that the visual system stores a *single* three-dimensional model for each object. If this were the case, the aligned model $T_{ij}(M_i)$ would additionally have to be projected onto two-dimensional space by some projection operator $\mathbf{P}$ before being compared to the image $X^{(2D)}$ projected onto the retina. Such a *full alignment* scheme would be object-centred and view-independent.

Alternatively, Ullman remarks that the brain may "store a number of models corresponding to sufficiently different viewing positions" (Ullman, 1989, p. 228) of the same

object. This method of representing objects is "*view-dependent* [emphasis added], since a number of different models of the same object from different viewing points will be used", but at the same time "it is expected (...) to be *view-insensitive* [emphasis added], since the differences between views are partially compensated by the alignment process" (Ullman, 1989, p. 228). Hence, the recognition performance of visual systems relying on view alignment methods should generally be *view-independent* (as long as (a) models of the viewed object are stored in memory and (b) the feature points necessary to access the corresponding model stored in memory are visible).

Linear Combination of Views. Another object recognition scheme that stores multiple views of the same object is *recognition by linear combination of two-dimensional views* (Ullman & Basri, 1989). This approach is based on the mathematical fact that a two-dimensional image obtained from the orthographic projection of a three-dimensional object can be expressed as a linear combination of a small number of two-dimensional model images of that same object. For example, Ullman and Basri (1989) have shown that objects with *sharp* edges undergoing rotation, translation, and scaling can be expressed as a linear combination of only four model images. Additionally, only three of these four images must be stored, as the fourth image can always be derived internally from the first three. Similarly, an object with *smooth* boundaries can always be expressed as a linear combination of six model images (Ullman & Basri, 1989).

Creating a new view of an object by linearly combining views stored in memory is done as follows: Objects are modelled by a 'set of object points', and it is assumed that "it is known which are the corresponding points in the different pictures" (Ullman & Basri, 1989, p. 11). A new view is thus obtained by linearly combining the coordinates of corresponding points according to the specified coefficients. The latter can separately be defined for the x- and y-coordinates.

For example, consider the apex of the pyramid depicted in Figure 1 and let $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ be its coordinates in the first, second, and third image of Figure 1 A, respectively. The apex' coordinates (x, y) in the new view illustrated in the left picture of Figure 1 B are given by $x = \alpha^i x_i$ and $y = \beta^j y_j$ with coeffi-

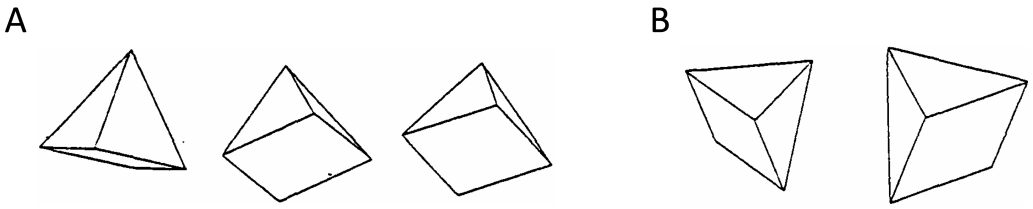

Figure 1. Example of the linear combination approach to object recognition proposed by Ullman and Basri (1989). *(A)* Three 2D-views of a pyramid. The second and third images are obtained by rotating the first image in space. *(B)* Two linear combinations of the pyramid model. The left image is produced by rotating the first image of (A) around all axes in three-dimensional space prior to orthographic projection. The image on the right has additionally undergone translation and scaling.

Note. Adapted from *Recognition by linear combination of models* (A. I. Memo No. 1152), by S. Ullman and R. Basri, 1989, p. 12, Artificial Intelligence Laboratory, Massachusetts Institute of Technology. Cambridge. Copyright 1989 by Massachusetts Institute of Technology.

cients $\alpha = (0.343, -2.618, 2.989, 0)$ and $\beta = (0.630, -2.533, 2.658, 0)$ (Ullman & Basri, 1989, p. 12). Note that the fourth coefficient, not used in the example above, corresponds to the fourth view that can always be derived from the stored three.

Under this object recognition scheme, recognition of a viewed object is declared for a model if its stored views M_i can be linearly combined to match the two-dimensional image $X^{(2D)}$ of the viewed object. Schematically, this can be expressed as

$$\left\| \left(\sum_i c_i M_i \right) - X^{(2D)} \right\| < \theta,$$

where $c_i = (\alpha_i, \beta_i)$ are the x- and y-coefficients of the stored views and θ is some threshold that must not be exceeded for recognition to be declared.

As for recognition performance, this view-dependent model should perform uniformly well on views that are part of the space of views spanned by the stored models. In contrast, the model should perform poorly on views belonging to a space orthogonal to the one spanned by the stored views (Bülthoff & Edelman, 1992). For example, if all the stored views of a pyramid are simply a rotation of the leftmost pyramid of Figure 1 A around the vertical axis, the model would perform poorly when recognising a pyramid that is obtained by rotating the pyramid around its horizontal axis.

View Approximation. Another view-dependent model built on representing an object by multiple two-dimensional views is *view approximation by regularisation networks* (Poggio & Edelman, 1990). In this approach, the view of an object is represented by the coordinates of a finite number of 'feature points' visible in the image. Addition-

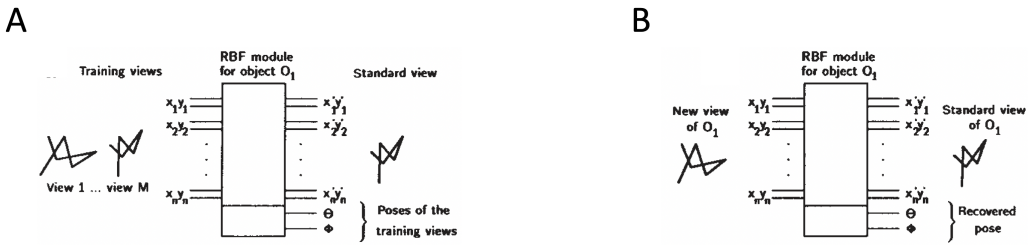

Figure 2. Illustration of the view approximation scheme by Poggio and Edelman (1990). *(A)* During training, the module is presented with several views of the same object and learns to produce the standard view of that object by transforming the coordinates of n 'feature points' visible in the image. *(B)* The module produces the standard view of the object and then compares this output with the stored standard view of the object in question. The object is recognised if the Euclidean distance between these two views does not exceed a certain threshold.

Note. Adapted from "A network that learns to recognize three-dimensional objects," by T. Poggio and S. Edelman, 1990, *Nature, 343*(6255), p. 264 (https://doi.org/10.1038/343263a0). Copyright 1990 by Nature Publishing Group.

ally, the approach is based on the following assumptions. First, it is assumed that for every object, there exists an object-specific smooth function that converts any view of the object into a fixed 'standard' view. Second, it should be possible to approximate this function from just a few different views of the object. Third, the application of a 'wrong' function (i.e., a function corresponding to a different object) is readily detectable (Poggio & Edelman, 1990).

Figure 2 schematically illustrates the process of object recognition based on view approximation: During training, the network is presented with varying views of an object (modelled by the coordinates (x_i, y_i) of some n feature points) and learns to recover the coordinates (x'_i, y'_i) of the standard view. This part is equivalent to approximating the smooth function that converts any view of the object into its standard view. The network applies the learned function to produce the standard view when presented with a novel view of the object (Fig. 2 B). Recognition is declared if the produced output matches the stored standard view.

In the recognition scheme proposed by Poggio and Edelman (1990), the approximation of the function that maps any view into the standard view is based on *generalised radial basis functions*. Essentially, a function $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ is called a *radial basis function centred at* ξ if its value on $\mathbf{x} \in \mathbb{R}^n$ depends only on the distance of $\mathbf{x}$ from ξ (Buhmann, 2003), hence the name 'radial'. Formally, this boils down to the existence of a function $\hat{\varphi}\colon [0, \infty) \to \mathbb{R}$ such that $\varphi(\mathbf{x}) = \hat{\varphi}(\|\mathbf{x} - \xi\|)$ for $\mathbf{x} \in \mathbb{R}$. In the context of object recog-

nition by view approximation, the radial basis functions can, for example, be chosen to be Gaussian. The output $\hat{\varphi}(\|\mathbf{x} - \xi\|)$ can be interpreted as the activity of one of the network's units that signals how close the input (i.e., a novel view of an object) is to one of the views learned during training. By combining the activities of all of its units, the network can recover the standard view of an object, which is then compared to the standard view stored in memory (Poggio & Edelman, 1990).

Being a viewer-centred model, view approximation is predicted to perform well on views close to those stored in memory. In contrast, as the novel views move further away from these stored views, the recognition performance should worsen gradually (Bülthoff & Edelman, 1992).

Experimental Evidence

The three classes of object recognition theories introduced above are predicted to yield conflicting patterns of recognition performance when presented with novel views of an object. Bülthoff and Edelman (1992) designed and conducted an experiment that takes advantage of this observation "to test the three theories [recognition by alignment, linear combination of views, and view approximation] directly in a psychophysical experiment involving computer-generated three-dimensional objects" (Bülthoff & Edelman, 1992, p. 60). Here, I will briefly explain the experimental setup, its underlying rationale, and the results obtained by the authors.

Methods. To test different classes of shape-based object recognition theories, Bülthoff and Edelman (1992) conducted a psychophysical experiment to test "the ability of human subjects (...) to generalise from familiar to unfamiliar views of visually novel objects" (Bülthoff & Edelman, 1992, p. 60). The stimuli used by Bülthoff and Edelman were computer-generated images of wire-frame objects, shown in Figure 3 A. The various viewpoints subjects viewed these stimuli from are illustrated by a viewing sphere centred at the recognition target (Fig. 3 B).

The experiment consisted of a training phase and a testing phase. During training, subjects were shown a "visually novel object (...) as a motion sequence of 2D views" (Bülthoff & Edelman, 1992, p. 61). These motion sequences spanned views from an azimuth of $75° \pm 15°$ and $0° \pm 15°$. The elevation was $0°$ for all of the views, and the motion sequence centred around the view

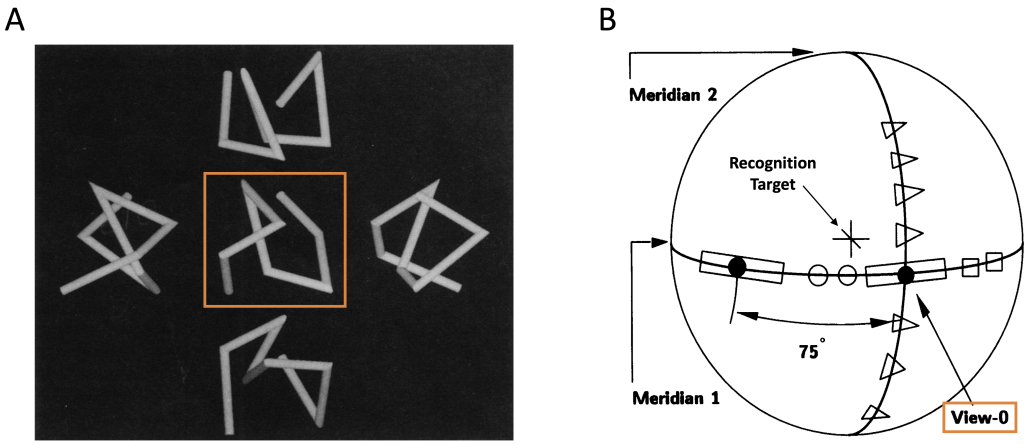

Figure 3. Experiment by Bülthoff and Edelman (1992) to test different classes of shape-based object recognition theories. *(A)* Paper-clip stimuli used in the recognition task. The image in the centre represents one view of a computer-generated stimulus. The remaining four images are obtained by rotating the centre image ±75° horizontally or vertically. *(B)* A viewing sphere indicates the various viewpoints that participants viewed the stimuli from. The sphere is centred at the recognition target.

Note. Adapted from "Psychophysical support for a two-dimensional view interpolation theory of object recognition," by H. H. Bülthoff and S. Edelman, 1992, *Proceedings of the National Academy of Sciences, 89*(1), pp. 61–62 (https://doi.org/10.1073/pnas.89.1.60). Copyright 1992 by the authors.

corresponding to an azimuth of 0° ('View-0' of Fig. 3 B) was always presented first to subjects.

Testing was done using a two-alternative forced choice paradigm. Participants were shown single views of an object, either target or distractor, and had to indicate whether or not the presented object was the target by pressing corresponding buttons. Additionally, subjects were instructed to do so "as quickly and accurately as possible" (Bülthoff & Edelman, 1992, p. 61). During the experiment, no feedback regarding the correctness of their answer was provided to participants.

The static views presented to subjects in the testing phase can be grouped into three categories (labelled 'INTER', 'EXTRA', and 'ORTHO' by the authors) based on their position relative to those presented during training. The INTER condition comprised views with an azimuth of 0° to 75°; these views were thus situated on the equator between those shown in the training phase. Views with an azimuth of 75° to 360° were grouped into the EXTRA category. Finally, the ORTHO condition consisted of views along the great circle orthogonal to the equator ('Meridian 2' of Fig. 3 B) (Bülthoff & Edelman, 1992). The three conditions are indicated by circles, squares, and triangles in Figure 3 B.

Bülthoff and Edelman intentionally designed the experiment this way because the three object recognition schemes introduced earlier predict different patterns of recog-

nition performance across the three conditions. Being a view-independent approach, a recognition scheme based on view alignment should perform almost perfectly in all three conditions. The linear combination of views theory also predicts excellent performance in the INTER and EXTRA conditions. However, this approach should perform poorly in the ORTHO condition, as views belonging to this category are part of a space orthogonal to that spanned by the views presented in training. Finally, the performance of the view approximation scheme is predicted to be highest for the INTER condition and lowest for the ORTHO condition, with performance in the EXTRA condition falling somewhere in between.

Results. The results of the experiment described in the previous section are shown in Figure 4 A: The error rate (false positives, in %) is plotted against the distance (in degrees) of the testing views from the reference view (i.e., straight on, see 'View-0' of Fig. 3 B). The mean error rate was lowest for the INTER (9.4%) and highest for the ORTHO condition (26.9%), with the error rate for the EXTRA condition (17.8%) sitting right in between (Bülthoff & Edelman, 1992). A statistical analysis revealed significant effects of condition ($F(2, 254) = 23.84$, $p < 0.0001$) and distance from the reference view ($F(6, 524) = 6.75$, $p < 0.0001$).

Qualitatively similar results were found for the same experiment with slightly different (more balanced) wire-frame stimuli (Fig. 4 B). Further, comparable results were obtained from an object recognition model based on view approximation in a simulated experiment (Fig. 4 C).

Additionally, the same experiment as in Figure 4 A was conducted with vertical and static training (results not shown). In the vertical training experiment, 'Meridian 1' and 'Meridian 2' of Figure 3 B were interchanged (i.e., the plane including INTER and EXTRA conditions was vertical, the ORTHO plane horizontal). In the static training experiment, single static images were presented during training instead of motion sequences. Notably, "subjects found it easier to generalize from a single familiar view in the horizontal plane than from an entire motion sequence within the vertical plane" (Bülthoff & Edelman, 1992, p. 62). This observation by the authors is based on the fact that error rates were significantly lower for both the INTER and EXTRA conditions in the static training experiment (6.9% and 27.3%) com-

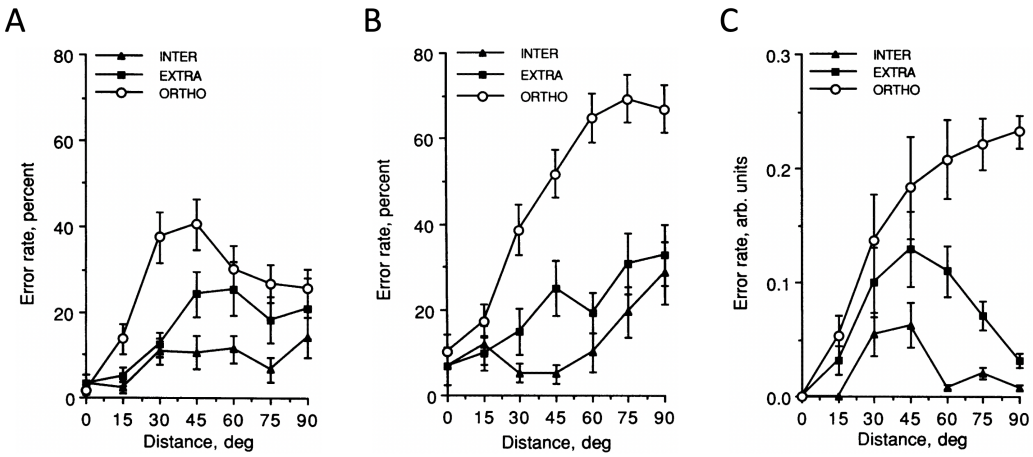

Figure 4. Results obtained by Bülthoff and Edelman (1992). *(A)* Type I error (false positives) plotted against the distance from the reference view ('View-0' of Fig. 3 B). *(B)* Same as in (A), except that more balanced stimuli were used in the experiment. *(C)* Same (simulated) experiment as in (B). Results represent the performance of a view approximation model.

Note. Adapted from "Psychophysical support for a two-dimensional view interpolation theory of object recognition," by H. H. Bülthoff and S. Edelman, 1992, *Proceedings of the National Academy of Sciences, 89*(1), pp. 62–63 (https://doi.org/10.1073/pnas.89.1.60). Copyright 1992 by the authors.

pared to the experiment employing vertical training (17.9% and 35.1%).

Discussion

According to the authors, the results of the experiments "fit most closely the prediction of the theories of the nonuniform 2D interpolation variety and contradict theories that involve 3D models" (Bülthoff & Edelman, 1992, p. 62). This conclusion is based on the observation that significant effects of condition were found in every experiment, contradicting the predicted near-perfect recognition performance of view-independent models that use single three-dimensional representations of objects, such as the recognition by alignment theory. The interpretation of the results provided by the authors is further supported by their ability to closely replicate the results of the experiments with human subjects in a simulated experiment using a 'prototype view approximation model'.

While Bülthoff and Edelman do admit that "it is possible that the subjects could not form the 3D representations required by the alignment theory" (Bülthoff & Edelman, 1992, p. 63), they point to another study (Edelman & Bülthoff, 1990) that yielded comparably poor results even though "training views were shown in motion and stereo" (Bülthoff & Edelman, 1992, p. 63). Also, the fact that the performance did not significantly drop in the static training experiment

renders the possible explanation mentioned at the beginning of this paragraph highly unlikely.

Finally, it should be considered that the experiments conducted by Bülthoff and Edelman (1992) used very specific wire-frame stimuli. These, of course, are highly dissimilar to most objects encountered daily, the latter often being a lot more symmetric than the stimuli used in the experiments of Bülthoff and Edelman.

References

Buhmann, M. D. (2003). *Radial basis functions: Theory and implementations*. Cambridge University Press. https://doi.org/10.1017/CBO9780511543241

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences*, *89*(1), 60–64. https://doi.org/10.1073/pnas.89.1.60

Edelman, S., & Bülthoff, H. H. (1990). *Viewpoint-specific representations in 3D object recognition* (A. I. Memo No. 1239). Artificial Intelligence Laboratory, Massachusets Institute of Technology. Cambridge.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, *343*(6255), 263–266. https://doi.org/10.1038/343263a0

Tarr, M. J., & Kriegman, D. J. (2001). What defines a view? *Vision Research*, *41*(15), 1981–2004. https://doi.org/10.1016/S0042-6989(01)00024-4

Ullman, S., & Basri, R. (1989). *Recognition by linear combination of models* (A. I. Memo No. 1152). Artificial Intelligence Laboratory, Massachusets Institute of Technology. Cambridge.

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, *32*(3), 193–254. https://doi.org/10.1016/0010-0277(89)90036-X