# Model-Based Temporal Object Verification Using Video

Baoxin Li, Rama Chellappa, *Fellow, IEEE*, Qinfen Zheng, and Sandor Z. Der

***Abstract*—An approach to model-based dynamic object verification and identification using video is proposed. From image sequences containing the moving object, we compute its motion trajectory. Then we estimate its three-dimensional (3-D) pose at each time step. Pose estimation is formulated as a search problem, with the search space constrained by the motion trajectory information of the moving object and assumptions about the scene structure. A generalized Hausdorff metric, which is more robust to noise and allows a confidence interpretation, is suggested for the matching procedure used for pose estimation as well as the identification and verification problem. The pose evolution curves are used to assist in the acceptance or rejection of an object hypothesis. The models are acquired from real image sequences of the objects. Edge maps are extracted and used for matching. Results are presented for both infrared and optical sequences containing moving objects involved in complex motions.**

***Index Terms*—Hausdorff matching, moving object recognition, object recognition, video processing.**

## I. Introduction

FOR many years, object recognition algorithms have been based on a single image or a few images acquired from different aspects. While advances have been made in simple constrained situations such as indoor environments, object recognition in natural scenes remains a challenging problem. Among the many difficulties, a prominent one is that in real applications, theoretically there exist infinitely many poses (orientations) for a given object. Therefore, two-dimensional (2-D) approaches, which are largely based on 2-D matching under some simplified transformation group, will not solve the three-dimensional (3-D) object recognition problem. To overcome the need for search in the viewpoint space, approaches based on geometric invariants have been proposed (for example, see [16] and [18]). Although the invariance approach is theoretically attractive, it would be difficult to apply it to complex objects in natural scenes. Appearance-based recognition schemes (for example, see [11]) try to tackle the viewpoint problem by using visual learning. In [11], the authors reported promising results for a test data set. Although appearance-based approaches do not require explicit feature extraction, their success relies on visual learning from a training set. A good training set is not always easy to obtain. Besides, due to shape variations, training images always contain some background region. Although when training, one can set the background to a uniform value (as in [11]), it is not always possible to black out the background at the recognition stage—one needs to know the object type and its exact orientation in order to do so, which is what a recognition algorithm is attempting to do. Backgrounds can greatly affect the projection of an input image onto the eigenspace. In addition, when the camera sensor is infrared, as in most surveillance applications, the object signature becomes too variable to be characterized by only a few images even at a fixed pose. In [9], some recognition algorithms including several learning algorithms were compared, using a large database containing over 17 000 images of ten object classes. It was reported that even the best recognition results were unsatisfactory for this infrared database. One possible explanation for the results in [9] is that when objects have abundant pose variations, the appearance manifolds become heavily overlapped, making recognition harder. In such a situation, one may have to resort to some geometric (shape) features, which, unfortunately, are again dependent on viewpoint.

An interesting observation is that when the object is moving, human beings can quickly guess its pose, and then verify some features unique to that pose. This suggests that additional information can be exploited to make object recognition more feasible when a video sequence is available. This paper presents a technique for model-based temporal object verification/identification. In a sense, verification and identification are constrained cases of recognition. To be specific, in this paper, identification refers to the following problem: given an image sequence containing a moving object, to identify the object as one of a few hypotheses; or, to identify the desired object in a sequence containing multiple objects. Identification is dynamic in that we have a time-evolving scene due to object motion and possible sensor motion. Verification is used in a slightly different situation, which answers the following questions: Is this the object seen in the previous frames? and How confident of this am I? This is especially interesting in situations of temporary loss of tracking due to, for example, occlusion by other objects. Verification is in a sense similar to the tracking problem but here it emphasizes the acceptance or rejection of a certain object hypothesis, rather than just tracking by using some features. Obviously, model-based verification/identification has many applications. For example, in visual autonomous surveillance as in following a face in the crowd, the recognition problem can often be reduced to the verification/identification problem.

Manuscript received March 12, 1999; revised February 26, 2001. This work was supported by the Advanced Sensors Consortium (ASC) sponsored by the U.S. Army Research Laboratory under the Federated Laboratory Program, Cooperative Agreement DAAL01-96-2-0001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael R. Frater.

B. Li is with Sharp Laboratories of America, Camas, WA 98607 USA (e-mail: bli@sharplabs.com).

R. Chellappa and Q. Zheng are with the Center for Automation Research University of Maryland, College Park, MD 20742 USA.

S. Z. Der is with the Army Research Laboratory, Adelphi, MD 20783 USA.

Publisher Item Identifier S 1057-7149(01)04479-7.

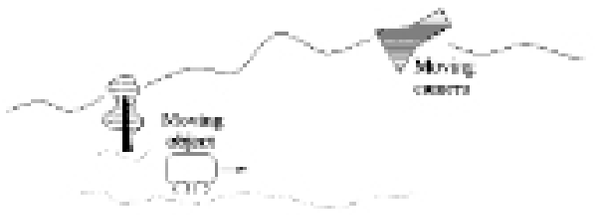

Fig. 1. Typical identification/verification setup using video from a moving camera platform.

In applications such as visual autonomous surveillance, the camera itself is often moving during the acquisition process. A general setup for this kind of problems is illustrated in Fig. 1. Due to camera motion, a sensor motion compensation process is often needed to remove the unwanted camera motion if we want to detect the object based on its motion.

In this paper, from image sequences containing the moving object, the 3-D pose of the object is estimated at each time step. Pose estimation is formulated as a search problem, with the search space strictly constrained by the motion trajectory information of the moving object and assumptions about the scene structure. A generalized $L_p$ version of the Hausdorff metric [1], which is more robust to noise and allows a confidence interpretation, is suggested for the search problem. The pose evolution curves are used to assist in the acceptance or rejection of an object hypothesis. Experiments on several sequences are presented. The experiments demonstrate how the concepts and algorithms for model-based temporal identification/verification could work in real applications.

## II. Matching Based on the Hausdorff Metric

The Hausdorff metric [7] is a mathematical measure for comparing two sets of points in terms of their least similar members. Formally, given two finite point sets $A = \{a_1, \ldots, a_p\}$ and $B = \{b_1, \ldots, b_q\}$, the Hausdorff metric is defined as

$$H(A, B) = \max\{h(A, B), h(B, A)\} \tag{1}$$

where

$$h(A, B) = \sup_{a \in A} \inf_{b \in B} \|a - b\| \tag{2}$$

and $\|\cdot\|$ is an underlying norm. If a model image and a scene image are first processed to give two characteristic point sets, then the model-scene matching is realized by comparing the point sets in terms of the Hausdorff metric. Intuitively, when there are multiple models, recognition is simply done by computing the corresponding Hausdorff distances between the models and the scene, and then picking out the best match.

### A. Some Modified Versions of the Hausdorff Metric

Although theoretically attractive, the Hausdorff metric $H$ is not directly usable in practice, because the sup or max operation in the definition makes $h$ and hence $H$ very sensitive to noise—a single noisy point can pull the value of $H$ far from its noise-free counterpart. Some modifications have therefore been proposed in the literature. For example, in [9], a weighted sum version was proposed and found to slightly improve the recognition rate; and in [5], a $K$th ranked partial "distance" $h(A, B)$ was used to detect a model in a static scene. The same partial "distance" was also used to track people in [8]. Although these modifications improve the robustness in practice, the obtained "distances" (a weighted one in [9] and a $K$th ranked one in [5]) no longer possess the properties of a metric. That is to say they are not real *distances* in the strict sense. We argue that being a metric (i.e., obeying the axiomatic rules for a metric) is important because when doing identification or verification, generally we have several hypotheses, and we need to use a measure that can reflect our confidence in choosing one over the others. This is not like detection or tracking, where one only needs to find an optimal match for a given mask. For example, it's easy to construct examples where a partial distance does not give a measure of similarity between point sets. Although these examples are unlikely to occur, one does face difficulties when the models are relatively simple point sets (with not too many points) while the scene is highly cluttered. Therefore, the above-mentioned modified versions of the Hausdorff distance do not necessarily offer good measures for comparison among different models.

### B. $L_p$ Version of the Hausdorff Metric

Another equivalent representation of the Hausdorff metric is (see [6])

$$H(A, B) = \sup_{x \in X} |\rho(x, A) - \rho(x, B)| \tag{3}$$

with

$$\rho(x, A) \triangleq \inf_{a \in A} \{\rho(x, a)\}$$

where $X$ is a set and $\rho$ a metric such that $(X, \rho)$ is a metric space, and $A \subseteq X$ and $B \subseteq X$. In the image analysis context, $X$ can simply be the set of all the image grid points, and $\rho$ is usually the $L_2$ norm, while $A$ and $B$ are two compact sets in the image plane. In this paper, we use edges as the features for matching; thus, $A$ and $B$ are just edge maps derived from intensity images.

To alleviate the instability in (3) due to the sup or max operation, Baddeley [1] has suggested an $L_p$ average as follows:

$$H^p(A, B) = \left[\frac{1}{n(X)} \sum_{x \in X} |\rho(x, A) - \rho(x, B)|^p\right]^{1/p} \tag{4}$$

where $n(X)$ is the number of points in $X$, and $1 \leq p < \infty$. So defined $H^p(A, B)$ is still a metric, and topologically equivalent to $H(A, B)$, but is more robust to noisy data since the contribution of a single point has been weighted. Also, by using the average, (4) has an "expected risk" interpretation: given $A$, a set $B$ which minimizes $H^p(A, B)$ is one which maximizes the pixelwise likelihood of $\{\rho(x, A) = \rho(x, B)\}$ (if $A$ and $B$ are treated as random sets). In applications, a cutoff function $w(t, c) = \min\{t, c\}$, for a fixed $c > 0$, is incorporated into (4) to give

$$H^p(A, B) = \left[\frac{1}{n(X)} \sum_{x \in X} |w(\rho(x, A), c) - w(\rho(x, B), c)|^p\right]^{1/p}. \tag{5}$$

The resulting $H^p(A, B)$ is again a metric, and topologically equivalent to $H(A, B)$. Note that in practice it is unnecessary to

compute $\rho(x, A)$ by its definition (i.e., by computing $\rho(x, y) = ||x - y||$), which is too expensive, especially with the $L_2$ norm. Instead, distance transformations [3] are used. Thus, using a supporting set $X$ will not cause significant extra computation, although $X$ is larger than $A$ and $B$.

### C. Identification/Verification with $H^p$

Given two point sets, $H^p$ provides a similarity measure between them. When this measure is applied to the identification/verification problem, we are concerned not only with how good the match is but also with where the match happens in the scene. It would be meaningless to compute $H^p$ between a small model and a large scene image. Instead, usually a region of interest (ROI) is detected first, and matching is carried out between the ROI and the model. In particular, in identification problems, given the edge map $R$ of an ROI from the scene image and $m$ models $M_i, i = 1, \ldots, m$, the task is to find a model $M_j$ and a transformation $T' \in \mathcal{T}$ such that

$$H^p(R, T'(M_j)) = \min_{i=1}^{m} \min_{T \in \mathcal{T}} H^p(R, T(M_i)) \qquad (6)$$

where $\mathcal{T}$ is an allowed transformation group for the application. Such $M_j$ will be regarded as the potential object appearing in the current scene. Since $H^p$ is a metric, we can also interpret the values $\min_{T \in \mathcal{T}} H^p(R, T(M_i)), i = 1, \ldots, m$ as a measure of confidence of choosing $M_i$ in the current frame. If $m = 1$, then the problem is reduced to detecting an object in the scene; in addition, if the model is extracted from earlier frames in the sequence, the problem reduces to one of tracking and verification.

It is not hard to search over $\mathcal{T}$ when $\mathcal{T}$ is the translation group. However it is difficult to consider other transformation groups such as affine. Even if we consider only rotation and scale, the search becomes a daunting task. The authors of [8] have proposed an efficient search scheme for rotation using the fact that the image takes value only on a digitized grid. In Section III-B, motion-based segmentation is used to minimize the need for search over the scale space.

## III. Model-Based Pose Estimation and Object Verification

In this section, we present an approach to pose estimation and verification based on matching using the $L_p$ version of the Hausdorff metric, with the motion trajectory information from motion analysis being used as a constraint to reduce the search space. The model acquisition step is discussed in Section III-A. Section III-B gives a brief overview of a framework for detection, tracking and segmentation of moving objects in video acquired by a moving platform. Pose estimation and object identification are discussed in Section III-C. Section III-D discusses methods for excluding clutter from the ROI. The pose evolution curve is defined in Section III-E. Section III-F discusses the interpretation of $H^p$ as a confidence measure, and a confidence figure is defined. Experimental results are presented in Section IV.

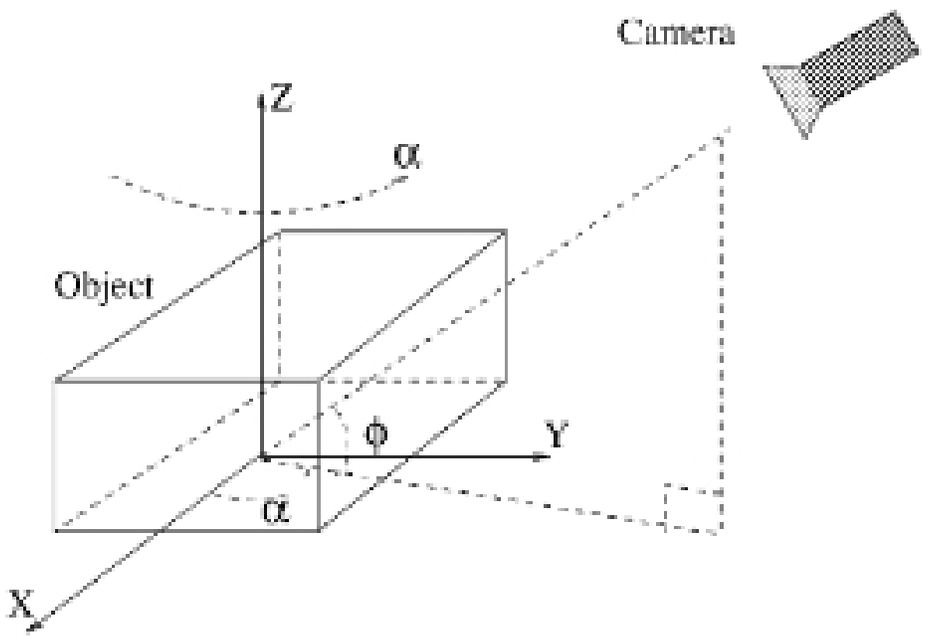


Fig. 2. Two angles defining the object orientation with respect to the camera under the assumption of level ground (i.e., the $X-Y$ plane is horizontal).

### A. Model Acquisition

When a 3-D object is subject to complex 3-D motion with respect to the camera, in general, multiple views of the object are needed for adequate modeling of the object. For a matching-based approach, images from these views constitute a model base. In general, there are two ways for constructing a model base: by using computer aided design (CAD) models or by extracting objects from real images. Three-dimensional CAD models allow one to easily manipulate the object orientation. However most objects of interest do not come with CAD models. In this paper, for the identification experiments, the models are constructed from real images: model images were taken at various camera depression angles, with the objects rotating horizontally. This allows the approach to extend to real applications easily: for any real object of interest, we can build its model by acquiring a set of images of the object at different viewpoints, hence relaxing the need for a 3-D CAD model.

Although in general, the orientation of a rigid object has three degrees of freedom, some assumptions can be made for specific applications. For example, if the object is on nearly level ground, as in most surveillance applications, its orientation can be characterized by only two variables. If we use an object-centered coordinate system, the object orientation is equivalent to the camera viewing angles, defined by two angles $\alpha$ and $\phi$ as illustrated in Fig. 2.

Notice that even under the above assumption, there are still infinitely many orientations in theory. But some observations can be made to determine the orientations that are characteristic. For example, with $\phi$ fixed, although $\alpha$ can vary from 0° to 360°, it is not necessary to store images at every degree of $\alpha$ since the object looks very similar when $\alpha$ changes only by a small number (say, less than 5°). A similar argument is valid for $\phi$, which takes values in the interval [0°, 90°]. More constraints can be included for a specific application. For example, in many applications, the value of $\phi$ can only change within a small range or can even be fixed. Research has shown that it seems that the human visual system represents objects only by a few 2-D views (e.g., [14]). Not much is known, however, about the number of views required for a specific object. In this work, we represent an object with a model base in which $\alpha$ and $\phi$ take on only a finite set of values.