DYNAMIC OBJECT IDENTIFICATION AND VERIFICATION USING VIDEO

B. Li, R. Chellappa and Q. Zheng

Center for Automation Research
University of Maryland
College Park, MD 20742-3275

S. Der

Army Research Laboratory
Adelphi, MD 20783-1197

ABSTRACT

In the paper, we introduce the concepts of dynamic object identification and verification using video. A generalized Hausdorff metric, which is more robust to noise and allows a confidence interpretation, is suggested for the identification and verification problem. Parameters from sensor motion compensation procedure are incorporated into the search step such that the Hausdorff metric based matching can be achieved efficiently under more complex transformation groups. An algorithm is proposed for identification/verification based on edge map matching using the generalized Hausdorff metric. Experiments on infrared video sequences are provided.

1. INTRODUCTION

For many years, object recognition algorithms have been based on a single image or a couple of images acquired from different aspects. While advances have been made for simple constrained situations such as indoor environment, object recognition in natural scenes remains a challenging problem. An interesting observation is that when given a video sequence of a moving object, additional information can be exploited and thus making object identification and verification more feasible. In applications such as visual autonomous surveillance, the camera itself is is often moving during the acquisition process. A typical setup of this kind of problems is illustrated in Fig. 1. Due to the motion of camera, object identification/verification should be carried out only *after* a sensor motion compensation process which removes the unwanted camera motion.

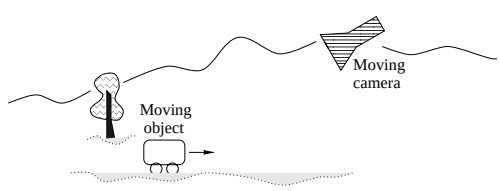

Figure 1: A typical setup of identification/verification using video

Dynamic identification and verification are the core tasks of the problem when a video sequence is available. We use identification and verification to mean different operations in this context. To be specific, dynamic identification refers to the following problem: given an image sequence containing the moving object, to positively identify the object type among a few hypotheses. Identification is dynamic in that we have a time-evolving scene due to both sensor and object motion. By using multiframe temporal information, the identification process has increasing confidence as time evolves. Dynamic 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 am I? This is especially interesting in the situation of temporary loss of tracking due to, for example, occlusion by a big tree or other man-made objects. Dynamic verification is in a sense similar to the tracking problem but here it emphasizes the verification/rejection of certain object, rather than just tracking on a few feature points or a region of interest.

In the paper, a generalized L_p version of the Hausdorff metric, which is more robust to noise and allows a confidence interpretation, is used for the dynamic identification/verification problems. An algorithm based on edge map matching using the generalized version of Hausdorff metric is then proposed. Experiments on infrared video sequences are presented. The experiments demonstrate how the concepts and the algorithms for dynamic identification/verification work in real applications.

2. MATCHING BASED ON AN L_P VERSION OF THE HAUSDORFF METRIC

Hausdorff metric is a mathematical measure for comparing two sets of points in terms of their least similar members. The metric is defined as the maximum of the minimum distances from all members of point set A to point set B. Formally, given two finite point sets $A = \{a_1, a_2, \cdots, a_p\}$ and $B = \{b_1, b_2, \cdots, b_q\}$, the Hausdorff distance is defined as

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

where

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

and $||.||$ is an underlying norm between two points.

2.1. Some Modified Versions of Hausdorff Distance

Although theoretically attractive, 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.

This work was prepared through collaborative participation in the Advanced Sensors Consortium (ASC) sponsored by the U.S. Army Research Lab under the Federated Laboratory Program, Cooperative Agreement DAAL01-96-2-0001.

Some modifications have been proposed in applications. For example, in [2], a weighted sum version was proposed and found to slightly improve the recognition rate; and in [3], 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 [4]. Although these modifications improve the robustness in practice, the obtained "distance" (a weighted one in [2] and a K-th ranked one in [3]) is no longer 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 metric) is important because when doing identification or verification, generally we have several hypotheses, and we need to use a measure that can reflect the confidence of choosing certain 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 really give a measure of similarity between point sets. Although those examples are unlikely to come out of an edge detector, 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 Hausdorff distance do not necessarily offer a good measure for comparison among different models.

2.2. An L_p Version of the Hausdorff Metric

By using Lipschitz inequality, another well-known representation of Hausdorff metric in Eqn. 1 can be obtained as (see [8])

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

where X is a set and ρ a metric such that (X, ρ) is a metric space, and $A \in X$ and $B \in X$. In the image analysis context, X is simply the set of all the image grid points, and ρ is usually the L_2 norm, while A and B are edge maps from intensity images.

To alleviate the unstableness in Eqn. 3 due to the sup operation, Baddeley has suggested an L_p average [1] 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, Eqn. 4 has an "expected risk" interpretation: given A, a set B which minimizes $H^p(A, B)$ is that 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 Eqn. 4 to give

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

The resulting $H^p(A, B)$ is again a metric, and topologically equivalent to $H(A, B)$.

2.3. Identification/Verification with H^p

Given two point sets, H^p provides a similarity measure between them. When this idea is applied to identification/verification problem, we are concerned with not only how good the match is but also 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 the 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, $i = 1, ..., m$, the task is to find a model M_j and a transformation $T \in \mathcal{T}$ such that

$$H^p(R, M_j) = \min_{i=1}^{m} \min_{T \in \mathcal{T}} H(R, M_i) \tag{6}$$

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

It is easy to do the search over $\mathcal{T}$ When T is the translation group. However it is hard to consider other transformation group such as affine. Even if we consider only rotation and scale, the search becomes a daunting task. In the next section, we give an approach which allows more complex transformation without requiring full search in the transformation space.

3. DYNAMIC IDENTIFICATION/VERIFICATION USING VIDEO ACQUIRED BY A MOVING PLATFORM

In applications such as surveillance, the imaging sensor is typically of infrared type in order to operate at night, and the video sequences usually suffer from high levels of egomotion and are often obtained in poor conditions including low light and low resolution, which make feature detection hard and unreliable. The sensor is typically far away from the objects, and the objects are usually small. Therefore only the contour is relatively reliable, and hence the internal edge pixels are generally not considered.

3.1. A Framework for Detecting, Tracking and Segmentation

In [5], an algorithmic framework was proposed which integrates image sequence stabilization, moving object detection and segmentation, object tracking approaches, and forms a front-end of the automatic target recognition system. Given a sequence, the segmentation step in [5] provides an ROI, which is based on the motion analysis of the moving object. Stabilization is based on an affine transformation to model the sensor motion [7]. That is, the transformation between pixels of frame k and frame $k + 1$ are defined by

$$\mathbf{P}_1 = \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix} \mathbf{P}_0 + \begin{pmatrix} T_x \\ T_y \end{pmatrix} \tag{7}$$

where $\mathbf{P}_0 = (x_0, y_0)^T$ and $\mathbf{P}_1 = (x_1, y_1)^T$ are pixels of frame k and frame $k + 1$ respectively.

Segmentation greatly facilitates matching: recall from Eqn.5, a supporting set X is needed for computing H^p. In practice the smaller X is, the less computation is needed. It is desired that X is the smallest region covering the potential object. Segmentation not only provides a small ROI as X but also greatly decreases the search region for the min operation in Eqn. 6.

3.2. Allowing More Complex Transformation Groups

In Eqn. 6, we need to search within a transformation group to find the best match. However, except for the translation group, searching in other transformation groups is generally not feasible. Based on the sensor motion compensation described in previous section, we can avoid searching for other transformations except translation. For example, when doing verification, we can first estimate the scaling from a model to the scene using the size of the ROI (this step, however, needs to account for the inaccuracy in the segmentation step), then use the affine parameters in Eqn. 7 to warp the models, followed by a search over translation group. A simpler example is when the sensor has fast looming motion towards the objects In this situation, without using affine transformation, we can assume that the object is subject to a scaling with the scale factor s estimated from the affine parameters as

$$s = \sqrt{\frac{r_{11}^2 + r_{12}^2 + r_{21}^2 + r_{22}^2}{2}} \tag{8}$$

3.3. Excluding Clutters In the ROI

The detected ROI contains not only the potential object but also background clutter. According to Eqn. 5, every edge pixels within the ROI will contribute to $H^p(A, B)$, which is unreasonable. When doing identification, the following technique is used to exclude clutters before calculating H^p: given a model M and an ROI R, we keep those points in ROI only if they are within certain distance of $T(M)$. Here $T(M)$ means the M has been subject to a transformation T. That is, a new ROI R' is formed by

$$R' = \{x : \forall x \in R \ \ and \ \ \rho(x, T(M)) < t\} \tag{9}$$

where t is a small positive number. On the other hand, if we are doing verification, the model is typically itself an ROI from previous frames. In this situation, the motion boundary estimated in [5] will be used as M in Eqn.9, and due to the inaccuracy of the boundary, a larger t should be used.

3.4. Interpreting H^p As A Confidence Measure

The nice properties of H^p allow a confidence interpretation. For the identification problem, this means at each step the H^p value for each model is treated as a measure of *confidence* in choosing a certain model: the smaller this number is, the more confident we are of choosing the model. If multiple models are kept as frames are processed, although at some time we may make the wrong choice, the future updates of the confidence level will hopefully provide the right choice.

For the verification problem, a confidence interpretation is also helpful: whenever we notice a sharp decrease in confidence, what may have happened is that the object is no longer the previous one, *or* the orientation of the object has changed dramatically. The information can be used to update the model hypotheses. Verification, in this regard, is similar to a tracking problem like in [4], with the following distinctions:

- the object is detected by motion analysis rather than specified by a human being;
- the camera is subject to motion;
- the object is small and can not provide dense edge information (thus only using a partial "distance" inevitably results in a lot of false matches).

4. EXPERIMENTAL RESULTS

We tested the algorithms on sequences acquired by an infrared camera mounted on a helicopter flying towards a tank. Due to the helicopter motion, the scaling becomes significant within a few frames. Fig. 2 illustrates the verification procedure. A moving object is first detected by the method proposed in [5], then an ROI is formed and processed to get the edge map. This edge map, used as the model, is verified in subsequent frames. In Fig. 2, the ROI from frame 302 is superimposed on frame 310, 315, and 320, respectively, after the locations have been estimated using Eqn. 6. Note the substantial scaling of the object (typically a scale factor 1.02 is obtained by Eqn. 8 for two consecutive frames). However, by compensating the scaling using Eqn. 8, the algorithm is able to locate the tank and report small H^p values (meaning high confidence). In experiments presented in this paper, p and c in Eqn. 5 are fixed as 1 and 4 respectively, and t in Eqn. 9 is 5. The edge maps were detected using Canny's algorithm [6].

Fig. 4 shows how identification works. Three hypotheses are provided, with the ground truth being model 2. One can see from the figure that, although in some frame the algorithm reports false identification result, overall the confidence of choosing the model 2 is higher than choosing the others. To see this more clearly, we plotted in Fig. 3 the H^P value for each model from frame 318 to 327. The overall confidence over model 2 is obvious.

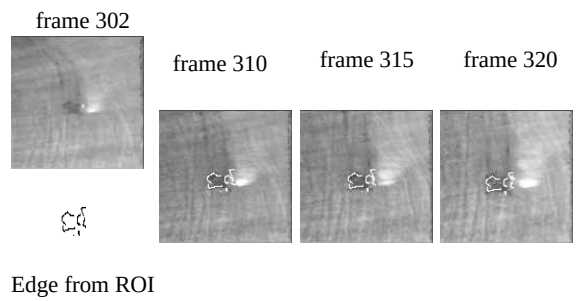

Figure 2: Dynamic verification with H^p metric: a moving object is first detected and its edge map is used in following frames for verification purpose.

5. CONCLUDING REMARKS

The experiments demonstrate how the concepts and algorithms for dynamic identification/verification work on the real video. It should be pointed out that it has been implicitly assumed in the above analysis that the motion is two-dimensional. For example, in the experiments, the models are planar silhouettes, and the motion is mainly translation and scale, although the approach can also handle affine group as long as the stabilization step provides the correct parameters.

Unfortunately, using the planar points model, the group of transformation has to be limited to planar transformation group. Generally speaking, the six-parameter affine transformation will not be able to bring a 2-D model into alignment with the scene if the 3-D motion induced by either the moving target or the sensor is so dramatic that no planar motion model is a good approximation. This situation is worsened if the frame rate is low while the sensor motion is fast, which means that the scene will suffer from larger 3-D

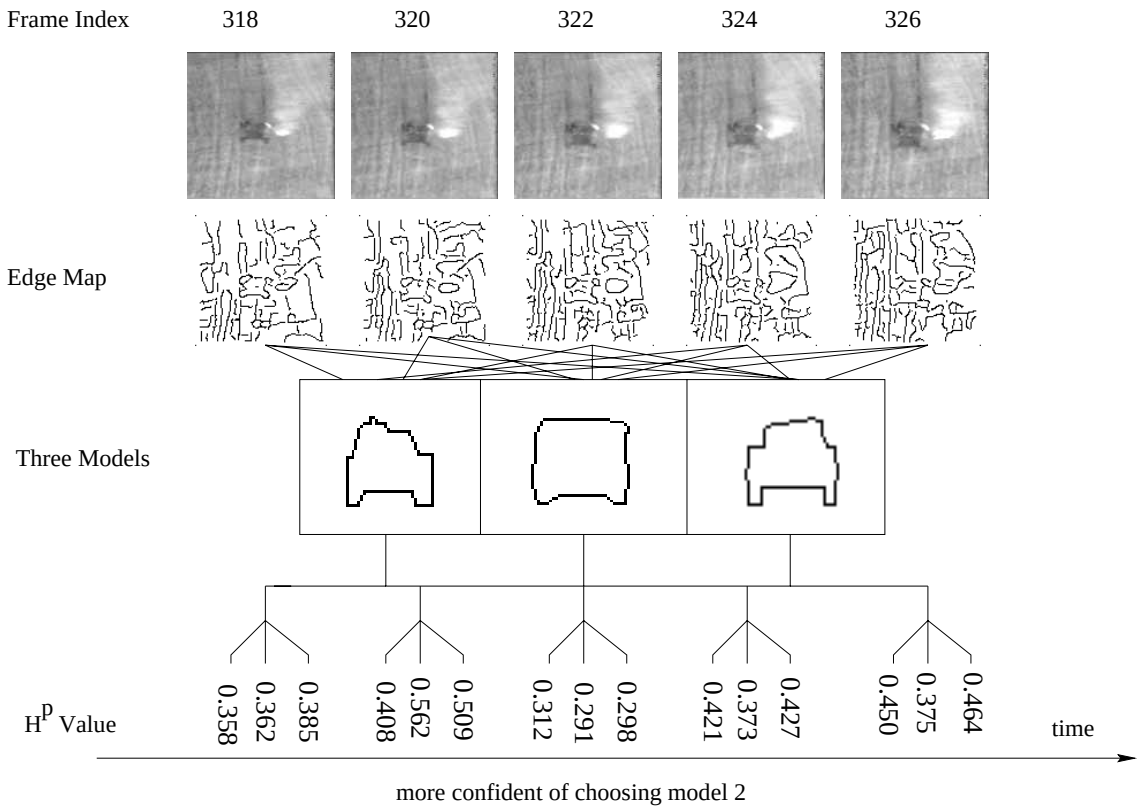

Figure 4: Dynamic identification with H^p metric: at each frame the models are compared against the detected ROI, and the H^p values are used to as a measure of the confidence of choosing certain model.

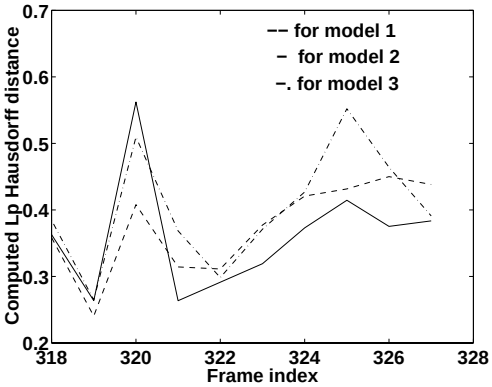

Figure 3: H^p values v.s. frame indexes : from frame 318 to 327.

changes within only a few frames. To handle this kind of situation, information about the relative pose between the camera and the object should be estimated so that a dynamic model can be generated after the confidence on current hypotheses become too low. The motion trajectory of the object readily provides certain constraints on the relative pose, and other information is still needed to get accurate pose estimates. We are currently working on this problem.

6. REFERENCES

[1] A.J. Baddeley, "Errors in binary images and an L_p version of the Hausdorff metric," Nieuw Archief voor Wiskunde, Vol. 10, pp.157-183, 1992.

[2] B. Li, Q. Zheng, S. Der and R. Chellappa, "Experimental Evaluation of Neural, Statistical and Model-Based Approaches to FLIR ATR ," Proc. of SPIE, Vol. 3371, April 1998.

[3] D. Doria and D. Huttenlocher, "Progress on the Fast Adaptive Target Detection Program," RSTA Technical Reports of the ARPA IU Program, pp. 589-594, 1996.

[4] D. Huttenlocher, J. Noh and W. Rucklidge, "Tracking Non-Rigid Objects in Complex Scenes," Proc. ICCV, 1993.

[5] B. Li, Q. Zheng, and S. Der, "Moving Object Detection and Tracking in FLIR Images Acquired by A Looming Platform," Proc. of JCIS, 1998.

[6] J. Canny, "A Computational Approach to Edge Detection," IEEE Trans. PAMI, Vol. 8, pp.679-698, 1986.

[7] S. Srinivasan and R. Chellappa, "Image Stabilization and Mosaicking Using the Overlapped Basis Optical Flow Field," Proc. of IEEE-ICIP,1997.

[8] H. Federer, "Geometric Measure Theory," Springer-Verlog, 1967.