Model Generation for Video-based Object Recognition

Humera Noor, Shahid H. Mirza
Faculty of Electrical & Computer Engg.
NED Univ. of Engg. and Tech.
Karachi - 75270, Pakistan
humera|deanece@neduet.edu.pk

Yaser Sheikh, Amit Jain, Mubarak Shah
School of Electrical Engg. & Computer Science
University of Central Florida
Orlando, FL 32816, USA
yaser|ajain|shah@cs.ucf.edu

ABSTRACT

This paper presents a novel approach to object recognition involving a sparse 2D model and matching using video. The model is generated on the basis of geometry and image measurables only. We first identify the underlying topological structure of an image dataset and represent it as a neighborhood graph. The graph is then refined by identifying redundant images and removing them using morphing. This gives a smaller dataset leading to reduced space requirements and faster matching. Finally we exploit motion continuity in video and extend our algorithm to perform matching based on video input and demonstrate that the results obtained using a video sequence are much robust than using a single image. Our approach is novel in that we do not require any knowledge of camera calibration or viewpoint while generating the model. We also do not assume any constraint on motion of object in test video other than following a smooth trajectory.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene analysis - object recognition, time varying imagery

General Terms

Algorithms, Design, Experimentation.

Keywords

Video-based object recognition, morphing

1. INTRODUCTION

With the advent of high performance systems and increased storage capacities, it has now become possible to work on huge amount of image-based data. This calls for techniques to extract information on the basis of the contents of the images without human intervention and identify the objects present in them. There are a variety of approaches explored for object recognition, like, CAD-based, appearance-based and shape-based methods; however, each approach has its own set of limitations [6]. In each of the techniques, a model is generated which is then compared with an image to identify the object being tested. However, object recognition from a single view may fail when there is much similarity among test objects or when the background clutter or partial occlusion masks features of the object. Zhou et al [2] utilized the temporal information present in video sequences for face recognition. They formulated a probabilistic model merging the dynamics and identity of humans obtained from video. However, they assumed certain constraints in the motion of persons while gathering their test data. Javed et al [3] presented a probabilistic framework for general object recognition using a video sequence containing different views of an object. They generated a model for each object in the training set capturing images at known viewing angles of camera and poses of objects.

In this paper, we present a novel strategy for object recognition. We use a set of reference images to generate an online sparse 2D model, estimate the underlying topological structure and, using image measurables only, arrange them in the form of a connectivity graph. We refine the graph using morphing, so as to remove the redundant images and finally use video matching for recognition of objects. The strength of our approach is that we don't need to know the object pose beforehand; and the video sequence could be shot over any arbitrary trajectory with objects following an unconstrained (but smooth) path. The use of video rather than a single image increases the confidence measure of the match.

The paper is organized as follows: Section 2 introduces how to identify the topological structure present in the images and arrange them as a neighborhood graph. It also describes the use of morphing for graph pruning. Section 3 highlights the use of the image database in conjunction with a test video sequence for automatic target recognition. In Section 4, we discuss the experiments conducted and the results obtained. Section 5 gives the conclusions and the future extensions.

2. MODEL GENERATION FOR OBJECTS

Any object can be modeled using either an object-centered or a view-centered representation [5]. The object-centered representations use the features from the objects, like boundary curves, surfaces etc, to describe the volumes of space. View-centered representations, on the other hand, depend

MM'06, October 23–27, 2006, Santa Barbara, California, USA.

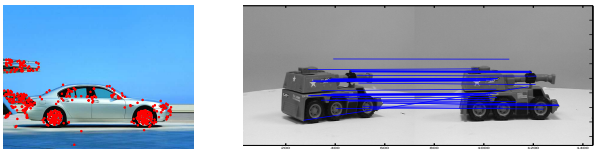

Figure 1: (Left)An image with SIFT points overlayed. (Right) Corresponding feature points for a pair of images.

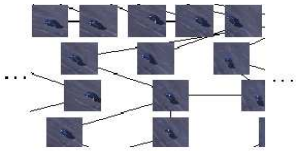

Figure 2: Portion of Neighborhood Graph for a car. Each image represents a node and edge is marked between pairs of neighbors.

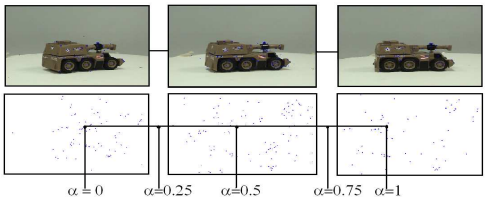

Figure 3: Bottom row identifies the feature points for images in top row. The feature points are checked for varying values of α to identify and remove redundant nodes.

on the outlook of objects from different viewpoints. These involve the use of aspect graphs and silhouettes for modelling. We have used the view-centered representation for generation of database, which makes the task of matching simpler. This is because the need for projection of model to 3D is no longer there and the features that are to be compared are in 2D [5].

The input to our database generation algorithm is a set of reference images, which have been arbitrarily extracted from a video sequence shot around an object. Our system tessellates the images around the viewing space of the object. The algorithm generates a neighborhood graph, where each image is identified as a node and the links between neighbors are specified as edges. The images are defined as neighbors on the basis of their proximity and extent to which they match with each other.

2.1 Development of Neighborhood Graph

Given a set of reference images, we propose a novel approach to tessellate them around the viewing space of the object while ensuring a minimal size of the database.

The algorithm begins by identifying the feature points in all the images of the repository. We have used the Scale-Invariant Feature Transform (SIFT) Operator to extract the distinctive features in the image. The features are invariant to image scale and rotation; and robust to changes in viewpoints and illumination. Feature correspondences are then identified using a fast nearest-neighbor algorithm [4], which are ultimately used to decide the presence or absence of linkage between nodes. Figure 1 shows SIFT points and matches identified for a pair of images.

For an image database having originally N images, an NxN link matrix is formed. A link between image pair (I_i, I_j) is marked if they are found to be neighbors.

The procedure for identification of neighbors is two-fold. In the first pass, we find the average Euclidean distance d for each image pair (I_i, I_j). For c corresponding points between two images, we have:

$$d(I_i, I_j) = \frac{\sum_{k=1}^{c} \sqrt{(I_{ik}^x - I_{jk}^x)^2 + (I_{ik}^y - I_{jk}^y)^2}}{c}. \quad (1)$$

For each image, the pair having minimum distance is selected as the neighbor, and an edge is marked between them. Considering this attribute as our seed point, we expand the region to include all those images in the neighborhood block, whose Euclidean distance falls within a certain threshold relative to the minimum value. This accounts for the out of plane images and handles arbitrary viewpoints. For an image database having originally N images, an NxN logical (binary) edge matrix is formed, where True in a cell represents a neighbor pair and vice versa. In the second pass, we apply the proximity constraint between successive video frames. This implies that two consecutive frames of a video sequence represent two images taken from adjacent (or closer) viewpoints and hence represent neighbors. Therefore:

$$Neighbor(I_i, I_{i+1}) = True. \quad \forall i \in Set\ of\ Frames \quad (2)$$

It may be noted that this second criterion improves the connectivity of the graph. In cases, where the image set is not from a true video sequence, and represents an arbitrary collection of images, only the first criterion would suffice. Figure 2 shows a portion of a graph that is generated for a car. Such a graph is generated for each object and stored as a model.

2.2 Multi-view Morphing

Once the Neighborhood graph is generated, it is refined using morphing. Seitz et al [7] introduced view morphing to generate novel views from varying viewpoints using only two images. Their approach is based on the principles of projective geometry, which can explicitly preserve 3D information. Given sparse correspondences between the image pair, view morphing works by rectifying the two images in such a manner that the corresponding points lie in the same scanline (a step known as pre-warping). This allows calculation of disparity map which helps in retrieving dense correspondences. Once the dense correspondences are known, the morph is generated using cross dissolve, and the resulting image is re-projected to its final position. Seitz's work could however be used to generate new views only along the line connecting the two original images. Later on Wexler et al [1] extended the concept to tri-view morphing and were able to synthesize morphs at any viewpoint within the boundaries of the triangle formed by the three images. Our graph pruning technique relies on the basic concept of morphing in that we synthesize the features for images following the same se-

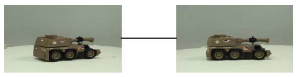

Figure 4: Updated neighborhood graph of Figure 3 after morphing. The central image was found to be redundant and thus removed.

quence as described above. However, we do not generate the complete image using morphing; rather we work on the sparse features identified by the feature detector. This saves us from computing the disparity map which takes time.

2.3 Graph Pruning

A view-centered approach leads to a space requirement that is larger than that of object-centered representation [5]. This is because many characteristic features are to be noted and there might be an overlap among the images. This requires special attention to be paid to keep the size of database at minimum.

We proceed by analyzing for each image if it represents a morph of its neighbors or not. To test any image I_i we begin with extracting its two adjacent images I_j and I_k and apply morphing on them to generate features and verify if they represent the features originally extracted from I_i or not. Given an image pair (I_j, I_k), with corresponding feature points p and q, we align the image pairs to have the corresponding points along corresponding scanlines and synthesize the features using Eq. 3 for varying values of α:

$$p_\alpha = p\alpha + (1-\alpha)q. \tag{3}$$

The features generated in this manner are compared with the original features extracted from I_i. For this, we have to iteratively engender and compare p_α for varying values of α. If there exists an α for which p_α represents the features of I_i, it means I_i could be generated using I_j and I_k and hence could be removed from the dataset. This procedure is demonstrated in Figure 3 and updated graph is shown in Figure 4. This proceeds till all the images in the database are exhausted. The same procedure is then repeated for images having larger number of neighbors.

After removing an image from the dataset, the neighborhood graph has to be updated, implying fixing the broken links that arise because of deletion of Image I_i. This involves creating new connections between original neighbors of I_i and estimating the distances between them. We accomplish this by forcing the neighbors of I_i to be neighbors. We know that: $Neighbor(I_i, I_j) = True \ \ and \ \ Neighbor(I_i, I_k) = True$. After removal of I_i, we have: $Neighbor(I_j, I_k) = True$. The Euclidean distance is updated as:

$$d(I_j, I_k) = d(I_j, I_i) + d(I_i, I_k).$$

3. VIDEO-BASED OBJECT RECOGNITION

One way to identify the target image is to generate a massive dataset of virtual views using morphing and compare the test image with all of them. This is inefficient and computationally expensive. We propose to initially match the test image with only those images stored in the database. This helps in identifying an approximate neighborhood of the image being examined. Once a seed image is found, the

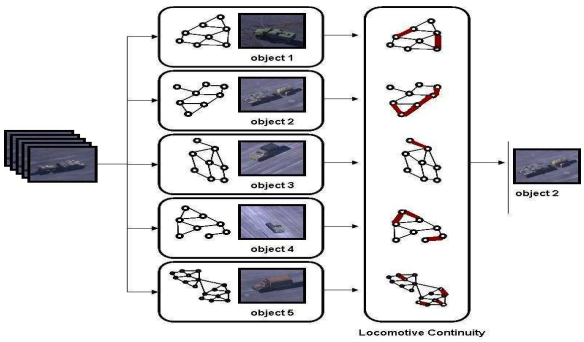

Figure 5: Using video sequence for matching. Each image of the test sequence is compared with each of the models and the matching links are highlighted. The smoothness of trajectory of the identified edges reflects the strength of matching.

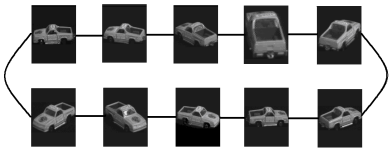

Figure 6: Linear graph generated for an object of COIL dataset.

virtual images around it could be generated using the morphing approach and compared with the test image.

In order to further strengthen the confidence measure of our detection results, we have used video sequences instead of single image for target recognition. The major advantage of this technique is that the video provides information of multiple views. Many objects in real world look alike, if observed from a particular viewpoint and completely different when observed from some other point of reference. Using a video for object recognition, we can exploit the fact that the two adjacent images in the video sequence represent proximally closer views of the object. Hence, the adjacent frames of the video sequence should point to the same (or adjacent) nodes of the neighborhood graph. Thus, a correct identification results in a smooth transition across the multiple nodes, following an unbroken trajectory in the model. On the other hand, an incorrect match results in jitters across the multiple frames, which helps in identifying the incorrect matches. Our approach for developing the topological structure of the images in database provides ease of traversing while using video sequence. As shown in Figure 5, given the stored networks of objects and a test video sequence, only the correct object follows a smooth trajectory along the graph and others suffer from discontinuities.

4. EXPERIMENTS AND DISCUSSION

To evaluate our approach for target recognition, we used the Columbia Object Image Library (COIL100) data set from the Columbia University and VIVID by DARPA. In COIL there are 72 images each of 100 objects. The viewing angle between these images is uniform, and this leads to a fairly linear neighborhood graph. See Figure 6 for neighborhood graph generated for a pick-up. In order to capture

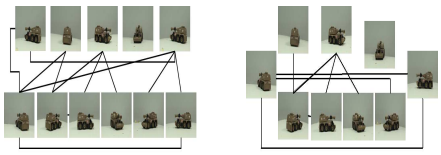

Figure 7: A portion of original neighborhood graph and its pruned network.

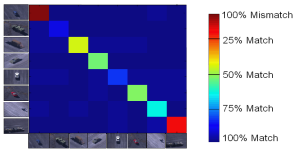

Figure 8: Single image matching resulting in a large number of incorrect matches.

the randomness of the real-world image capture, we shot our own video sequences following arbitrary trajectories.Figure 7 shows a portion of the neighborhood graph of one of the objects and the updated Network. Experiments show that our algorithm could generate the neighborhood graph with a precision of 97.86%. The system was able to reduce the image base to about 60% of its original size. Figure 8 shows results of single image matching for 8 objects of the VIVID data set, where each one is selected for matching against all the rest. The dark blue cells represent image match and the red cells identify mismatch. As can be seen, a lot of incorrect matches have been highlighted in case a single image matching technique is employed.

Video based matching provided significant improvement over single image matching. Single image matching gave 40% correct matches, while video-based recognition gave about 80% correct matches. The reason for the 20% incorrect matches is the high similarity of different objects at certain poses, which further increases the viability of our approach of using videos instead of single image for matching. The Figure 9 shows a smooth trajectory for the correct identification of motor bike. Figure 10 identifies an incorrect matching of a Humvee with the green truck by pointing discontinuities.

5. CONCLUSION

In this paper, we present an approach for video-based object recognition, which is a very crucial activity for content-based image retrieval. We generate an online sparse 2D model, based on geometry and image measurables only. Our system does not require camera calibration or prior knowledge of object pose. It does not assume a known 3D CAD model and does not place any constraints on motion of objects while video capture. Moreover, our approach results in a database of optimal size because of the removal of redundant images through morphing. We've shown that using our algorithm, the dataset size could be roughly reduced to 60%. We've used video sequences, instead of images, for object recognition and shown that this approach can efficiently remove false positives. Our approach for developing the topological structure of the images in database provides ease of traversing while using video sequence.

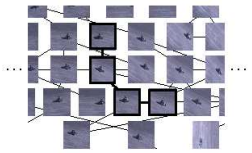

Figure 9: Smooth trajectory through the neighborhood graph for a correct match for motor cycle. The dark bounding boxes represent the images identified by the algorithm.

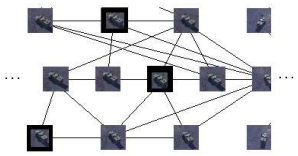

Figure 10: Jitters across neighborhood graph representing an incorrect match for a green truck.The dark bounding boxes represent the images identified by the algorithm.

One of the major strengths of our approach lies in the flexibility of framework. The neighborhood graph is adaptable; hence the model could be updated online during testing. Any new image could be linked to the network to provide additional information. Moreover, we work on the distinctive features of the image. This makes our approach extensible to IR images as well. In future, we intend to demonstrate the extensibility of our system. Moreover, the graph pruning could be further improved by using a more principled approach to provide higher reduction rate while reducing the time for pruning.

6. REFERENCES

[1] Y. Wexler and A. Shashua. On the synthesis of dynamic scenes from reference views. IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, SC, 2000, 576 - 581.

[2] S. Zhou, V. Krueger, and R. Chellappa. Face Recognition from Video: A CONDENSATION Approach. Proceedings of International Conference on Automatic Face and Gesture Recognition, Washington D.C., USA, 2002.

[3] O. Javed, M. Shah and D. Comaniciu, A Probabilistic Framework for Object Recognition in Video, Proceedings of International Conference on Image Processing, Singapore, 2004.

[4] David G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, 60, 2 (Nov 2004), 91 - 110.

[5] A. R. Pope, Model-Based Object Recognition - A Survey of Recent Research, Technical Report TR-94-04, University of British Columbia, January 1994.

[6] J.Xiao and M. Shah, Automatic Target Recognition Using Multi-View Morphing. Proceedings of SPIE on Automatic Target Recognition XIV, 2004, 391-399.

[7] S. Seitz and C. Dyer, View Morphing, Proceedings of ACM SIGGRAPH, 1996, 21 - 30.