Automatic Target Recognition Using Multi-View Morphing

Jiangjian Xiao Mubarak Shah

Computer Vision Lab, School of Electrical Engineering and Computer Science
University of Central Florida, Orlando, Florida 32816, USA
{jxiao, shah}@cs.ucf.edu

Abstract

This paper describes a novel approach to automatically recognize the target based on a view morphing database constructed by our multi-view morphing algorithm. Instead of using single reference image, a set of images or a video sequence is used to construct the reference database, where these images are re-organized by a triangulation of viewing sphere. At the vertex of each triangle, one image is stored in the database as the reference view from a specific viewing direction. For each triangle, our tri-view morphing algorithm can synthesize a high quality image for an arbitrary novel viewpoint amongst three neighboring reference images, and the barycentric blending scheme guarantees the seamless transitions between each neighboring triangles. Using the synthesized images, we apply appearance based recognition technique to recognize the target. In addition, using the proposed method, the pose of the object or camera motion can be approximately estimated. Several examples are demonstrated in the experiments to show that our approach is effective and promising.

1 Introduction

Automatic Target Recognition (ATR) deals with detection, recognition, and classification of targets from imagery. Popular methods of ATR include: CAD-based, appearance-based, and shape-based methods.

In CAD-based ATR, an explicit 3-D model of a target is generated and subsequently used in target recognition employing imagery acquired by a variety of sensors. The main step of CAD-based ATR is to estimate the pose of the CAD model so that the projection of 3D model matches with the queried image. Another popular approach is appearance based method [5, 4], where the 2-D intensity template of 3D target acquired from different viewpoints are stored as a model. The match between the model view and the input image is performed using simple correlation kind of matching. In shape based recognition [1, 2, 3], the contour or silhouette of the object is extracted, then the body silhouette templates are used to match the extracted shape. Efros et. al. presented an approach to recognize human actions at a large distance [3]. In their paper, they first recorded a large number of video clips to make a database. After characterizing the human motion parameters of an input video, they computed a spatio-temporal cross correlation to determine the most similar motion descriptor in the database.

However, it is not easy to acquire CAD models of all possible targets in the CAD-based approaches. Manual modelling of various targets is very burdensome and tedious for the model preparation stage. Currently, most of ATR approaches employ only single image, which can only provide the limited information about the object. If the pose or illumination of target is changed, these algorithms usually cannot give the correct solution. Therefore, the recognition using video (multiple images) is a good alternative, which has a potential to provide robust results

under variation of pose and illumination. In multiple images recognition, a set of images is taken form different viewing angles around the object, which provide more evidence compared to the single image recognition. These views can then be compared with the original images of an object model, and the appearance evidence can be accumulated. Employing the temporal information embedded in the video sequence, e.g., pose evolution curves, can assist in removing outliers and efficiently increase the confidence of recognition even under the variation of pose.

Seitz and Dyer proposed view morphing technique to synthesize a novel view given two reference images [7, 6]. In this framework, the novel view are interpolated according to the perspective geometry principles, which can provide more realistic results when blending the images of the scene. Xiao and Shah extended this framework into tri-view and multiple view morphing, such that the novel view can be generated in 2D space around the scene [11, 10]. VanMaasdam and J. Riddle proposed that the use of multiple range data acquired from different viewing angles to construct an integrated scene to perform ATR [8].

In this paper, we propose a novel approach to identify a target using view morphing database generated by employing a few reference images. These reference images can be a video sequence or sparse images captured by multiple cameras from different orientations and locations. Our approach is based on Multi-view morphing technique, which can synthesize a series of photo realistic virtual images using three or more images [10]. Based on the appearance based recognition technique, we compare the input images with the synthesized images to recognize the object. At the same time, the pose of the object or camera motion can approximately be estimated.

The paper is organized as follows. Section 2 introduces how to organize the reference images and build the view morphing database using our multiple view morphing algorithm. Section 3 describe two different schemes by using single input image and video sequence to automatically identify the object. In Section 4, we demonstrate how to evaluate our algorithm by using the COIL-100 image database, and also present our results.

2 Constructing Multiple View Morphing Database

Before recognizing the object for an input image, we build a view morphing database to store the key information of a set of objects. The input of our multiple view morphing algorithm is the reference images, which can be a video sequence or sparse uncalibrated images. Our view morphing database algorithm includes the following major steps:

First, using a two-stage wide baseline matching algorithm method employing edge corners [9], a number of corresponding points are automatically recovered to compute the fundamental matrix and epipolar geometry for each pair of reference images.

Second, for each of three neighboring reference images, a unique trifocal plane E and trifocal tensor are determined. The related viewing angle between each pair of images can also be computed.

Third, based on the viewing angles of reference images, the viewing sphere around the object is organized into an approximately uniform triangle tessellation as shown in Figure 1.

Fourth, using tri-view morphing algorithm, the novel synthetic view is generated according to the viewpoint position, which can be used to identify the input image using appearance based recognition approach.

2.1 Tessellation of Viewing Space

Given a set of reference images, we propose an approach to tessellate the viewing space to organize images and maintain the view morphing database at a minimal size.

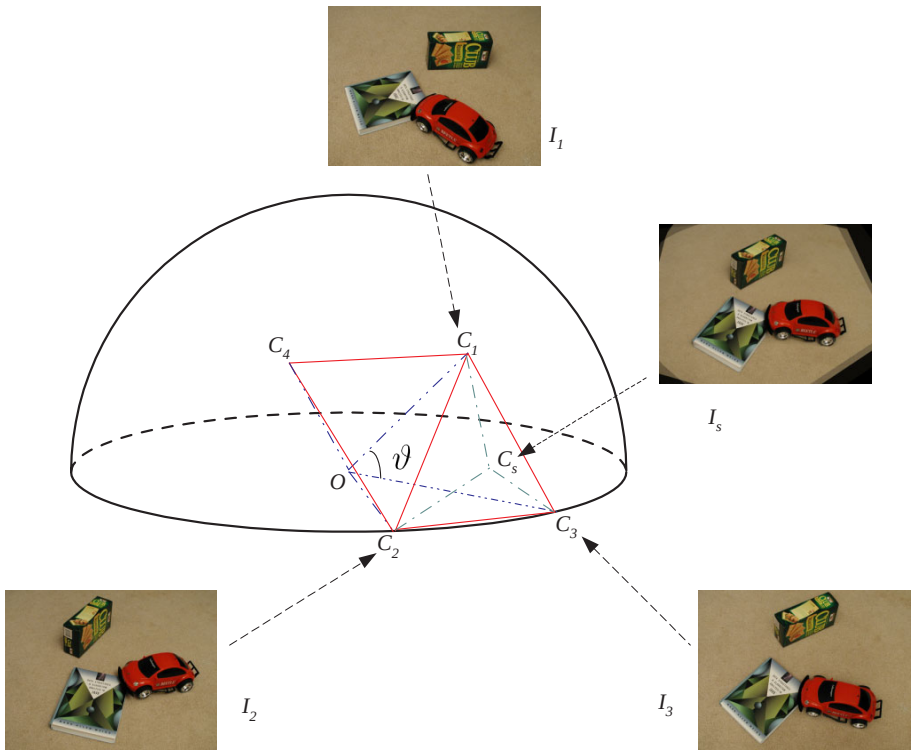


Figure 1: Given a set of reference images, a semi-sphere around the object can be approximately tessellated into a set of uniform triangles (red lines), where the object is located at the sphere center O. C_i is a triangle vertex, and blue lines, $\overline{C_i O}$, represent the viewing direction of camera i. ϑ is the viewing angle between neighboring view direction. I_i are reference images corresponding to camera i. I_s is a synthesized image generated by a camera located at c_s using the reference images I_1, I_2, and I_3.

Using Voronoi diagram, the viewing sphere around the object can be approximately divided into uniform triangles. Each angle between neighoring view directions is approximately equivalent. In our experiment, we have determined that the best angle, ϑ_o between two neighboring view directions is round $20 \sim 30^\circ$ (Figure 1). If the angle is larger, the quality of the disparity decreases due to the occlusion and specular reflections. If the angle is smaller, the database needs to store more images and corresponding disparity maps, therefore the size of the view morphing database becomes quite larger.

For each such triangle, the tri-view morphing algorithm can generate the high quality synthesized images, I_s, as shown in Figure 1. In order to obtain a seamless transitions between each neighboring triangles, we also apply the barycentric blending scheme to render the novel images.

As a result, given a set of ideal reference images covering a semi-sphere, the perfect tessellation with $\vartheta_o = 30^\circ$ requires 29 reference images, which are uniformly distributed on this semi-sphere. Based on this triangle tessellation of viewing space, any arbitrary novel view direction from the upper semi-sphere can be generated by using our tri-view morphing algorithm.

However, the original reference images may not be located at the perfect position, and some of them may be too close. In order to balance the database, we propose to use an average angle, $\bar{\vartheta}$, to measure the image redundancy, where $\bar{\vartheta}$ is the average viewing angle between a vertex and its neighboring vertices. If the average angle of one vertex is less than ϑ_{min}, we remove this vertex from database. Using this approach, we can efficiently control

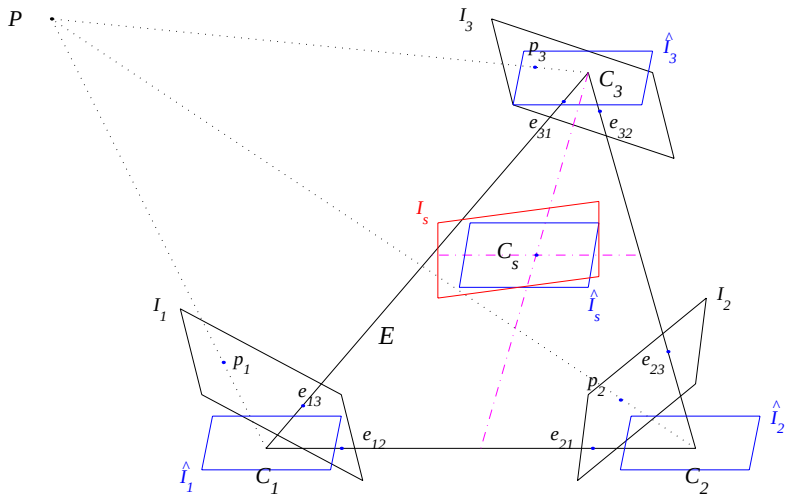


Figure 2: Tri-view morphing procedure. After automatically determining a focal plane E, which is constructed by three camera centers C_1, C_2, and C_3, three original images I_1, I_2, and I_3 are warped into parallel views $\hat{I}_1$, $\hat{I}_2$, and $\hat{I}_3$. The morphing image, $\hat{I}_s$, is blended by using the rectified images with correct disparity maps. The final image I_s at C_s is postwarped from $\hat{I}_s$.

the database size of the implicit model, where the upper bound of the database size is approximately determined. Once arriving the upper bound, our approach can provide an optimal balance between the database size and the quality of the synthesized view in the whole upper semi-sphere.

2.2 Tri-view Morphing

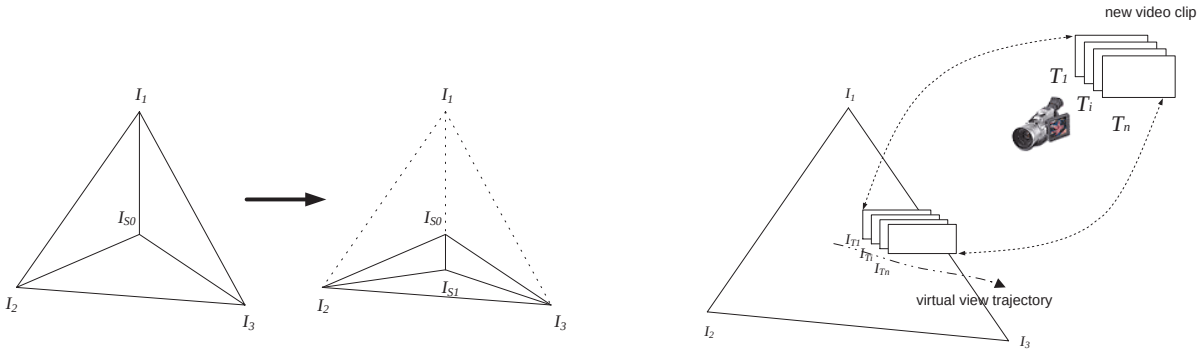
After triangle tessellation of a viewing space, we obtain a set of triangles. For each triangle, we first compute reliable corresponding points between each pair of images. Based on these points, a unique trifocal plane E is determined using their epipoles e_{ij} ($i, j \in \{1, 2, 3\}$ and $i \neq j$) as shown in Figure 2, where in camera C_1, the plane normal $N_{E1} = e_{12} \times e_{13}$; in camera C_2, $N_{E2} = e_{23} \times e_{21}$; and in camera C_3, $N_{E3} = e_{31} \times e_{32}$.

After the trifocal plane is determined, the three original images are warped into parallel views using the prewarping algorithm and employing the computed N_{E1}, N_{E2}, and N_{E3}. Also, the epipoles are projected into infinity. As a result of this warping, all epipolar lines in the three rectified images are pairwisely parallel. Next, for each pair of rectified images, corresponding epipolar lines are rotated to make them parallel to scanline directions.

Next, we use a pixel-to-pixel dynamic-scanline algorithm which uses an inter-scanline penalty to compute a dense disparity map for each corresponding rectified pair.

Finally, a barycentric tri-view blending function is determined according to the viewpoint position. Following the perspective geometrical principles, the morphing image, $\hat{I}_s$, is obtained by combining the blending function with the disparity maps. Then, a 5-point postwarping scheme is used to project the morphing image to a proper final position [11]. Figure 3 shows the tri-view morphing results for three “car-book” images.

Figure 3: A typical tri-view morphing scenario. The first three images are uncalibrated wide baseline reference images. The last three images are synthesized virtual views.



(a) (b)

Figure 4: (a) Estimation of closed view position for a single image using triple tree search. (b) Estimation of closed view trajectory for one video clip. If the whole video clip matches with some virtual view trajectory, the probability for recognition will increase.

3 Automatic Target Recognition Using Single Image and Video Sequence

In order to match a new image of the target, one simple way is to generate a large database of the virtual views, which record the appearance of the object from each possible viewing angle. Next, appearance matching can be used to compare the new image with the database to obtain the confidence of the identification. However, in this approach a very large image database has to be built to record all virtual views of the object, which need large space and time. Therefore, in context of multi-view morphing approach, we propose another searching scheme, "triple tree search", to reduce the time and space for the appearance matching as shown in Figure 4.a.

Assuming our implicit model for each object at least consists of three reference images, we propose to proceed as follows. First, we compare the novel image with the reference images I_1, I_2, and I_3, and obtain the probability of matching p_1, p_2, and p_3 (see Figure 4.a). If p_1 is less than p_2 and p_3, the closed virtual view will be located on the sub-triangle $I_2 I_3 I_{S0}$, where I_{S0} is the central virtual view amongst I_1, I_2, and I_3. Next, we can repeat the procedure, and determine the new sub-triangle of the closed virtual view. After N splits, the closed virtual view can be obtained, and the highest probability of the appearance matching can be computed efficiently. A computational analysis of the triple tree search can be described as follows: At each split, $\frac{2}{3}$ of the search area is eliminated and we are considering only $\frac{1}{3}$ of the area of triangle. By the N^{th} split, the area we consider drops exponentially to $(\frac{1}{3})^N$. Thus the number of iterations to be considered is usually very low, $N < 5$. Compared to the binary search, triple tree search is more efficient, i.e. in binary search by the N^{th} only $(\frac{1}{2})^N \gg (\frac{1}{3})^N$ of possibilities are eliminated.

Moreover, a sequence of images can also be used to robustly confirm the id of the target. The process for matching a sequence with the implicit model is similar to the single image matching shown in Figure 4.b. After

Figure 5: The images of 8 different cars in COIL100.

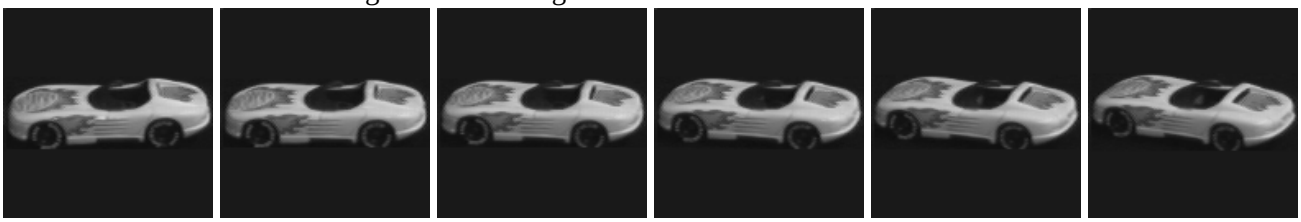


Figure 6: The image sequence of a toy car in COIL100 from 0° to 25°. The viewing angle interval between the consecutive frames is 0°.

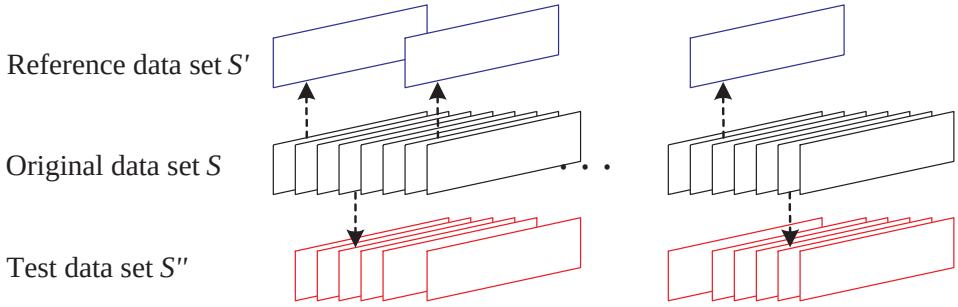


Figure 7: Evaluation using COIL100 database. The middle row is the original data set S. The top row is the reference data set S'. The bottom row is the test data set S''.

we can locate the match for the first frame of the novel sequence, the subsequent frames can be matched by searching in a local neighborhood of previous frames. Since the motion of the object or camera is continuous, the pose of the image should be also continuous, which provide a very effective temporal measurement. Therefore, if the estimated pose for one queried image set is smoothly changed from one to the successive one, the object can be classified into this category. If the curve suffers from a random jitter between adjacent frames, the recognition confidence for this set of reference image is low, which can be considered as the outlier.

However, in previous ATR approach using multiple images, if the gap between two neighboring reference images is large, the comparison between the input images (or video) with the sparse reference images may not generate a smooth temporal curve. In our multi-view morphing scheme, the gaps between neighboring reference images can be seamlessly filled by using image interpolation, and the novel image from any viewing angle can be pre-generated. Using these synthesized novel views, the accuracy of the pose temporal curve can be improved significantly.

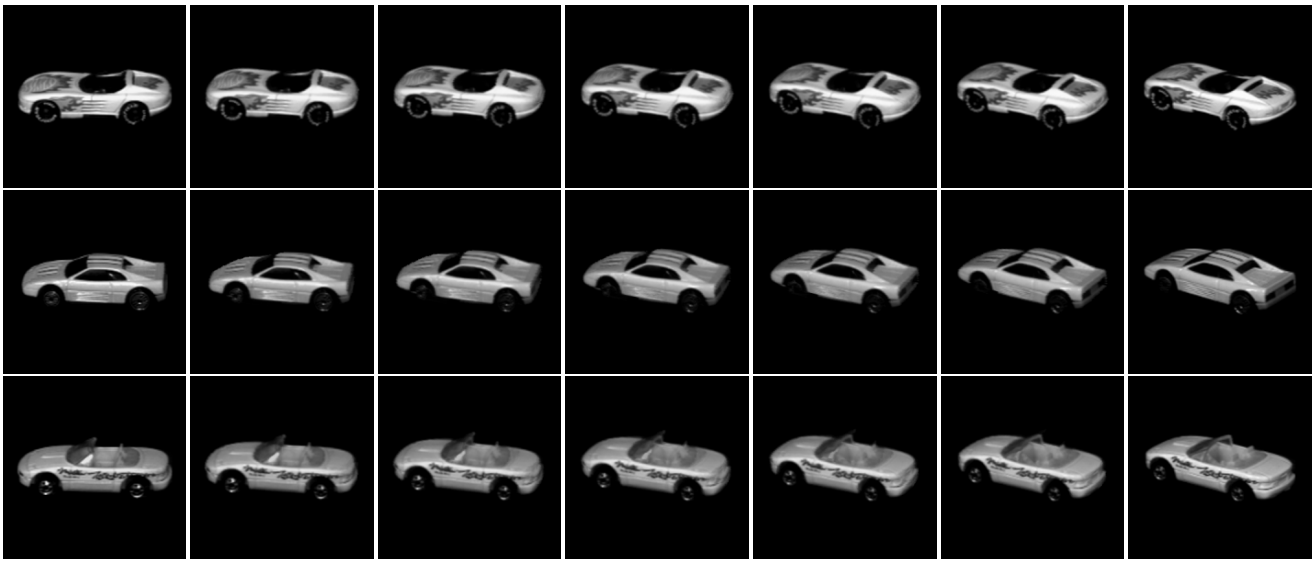


Figure 8: Three sequences of synthetic images generated by the view morphing database. The view angle between each pair of neighboring images is 5°.

4 Experiments

To evaluate our approach for ATR, we have used the data set from Columbia University to generate our implicit 3D database for object recognition. In Columbia object image library (COIL100), there are 100 objects, each of which has 72 pictures. However, in this data set, most of objects are very different, and it is easy to recognize. With the purpose of increasing the evaluation confidence, we selected 8 similar objects (toy cars) from the data set as shown in Figure 5. Furthermore, we have used the gray images instead of the color images to increase difficulty of recognition since the toys may have very distinguished colors. Unfortunately, in COIL100, no video is provided for this kind of moving objects. Therefore, we can only use this data set to test the single image recognition.

For each object, all images of an object are taken by a camera around the object from 0° to 360° in a 5° rotation interval. We denote these images as the original data set, S. Figure 6 shows some frames in a original car data set from COIL100.

Using our multi-view morphing framework, a typical tessellation with 30° requires 29 reference images uniformly distributed on an upper semi-sphere. Since the camera's pitch angle is fixed in the COIL100, we only select 12 images as the reference images with a 30° rotation interval, which are denoted as reference data set S'. The remaining 60 images in the data set are test data set S'', which is used for input to test our algorithm as shown in Figure 7. Using this twelve reference images, we first apply multiple view morphing algorithm to generate the view morphing database for each of object, which store the reference images and disparity maps between the neighboring image pairs. Based on the view morphing database, the view at any arbitrary location can be synthesized as shown in Figure 8.

Figure 9 also checks the correlation error between the real images (ground truth) with corresponding synthetic images generated at the same viewing direction. Since the synthetic images may have different scales compared to the real images, we apply an affine transformation to warp the synthetic images and then check the Sum of Absolute Differences (SAD). The SAD values are zero when the synthetic images are coincident with the reference images acquired at 0°, 30°,60°, etc. For the other viewing angles, there are very small differences

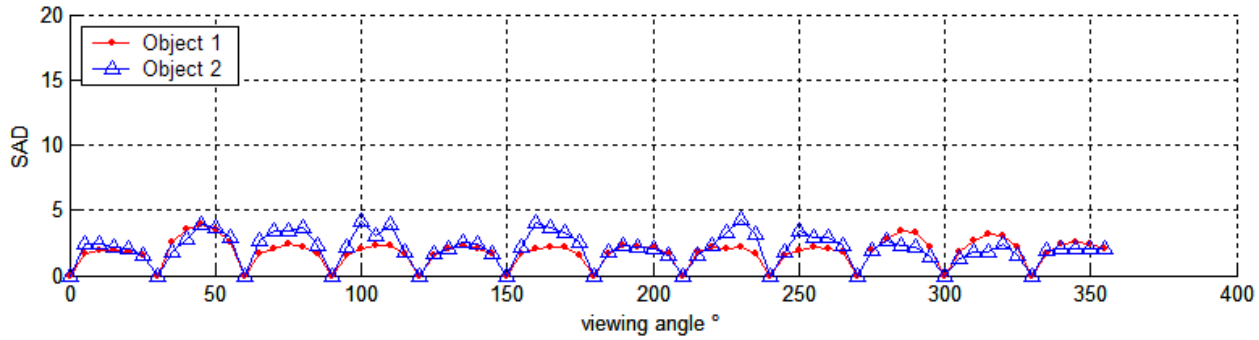


Figure 9: The Sum of Absolute Differences (SAD) between real images and corresponding synthetic images. The differences between the corresponding images are very small.

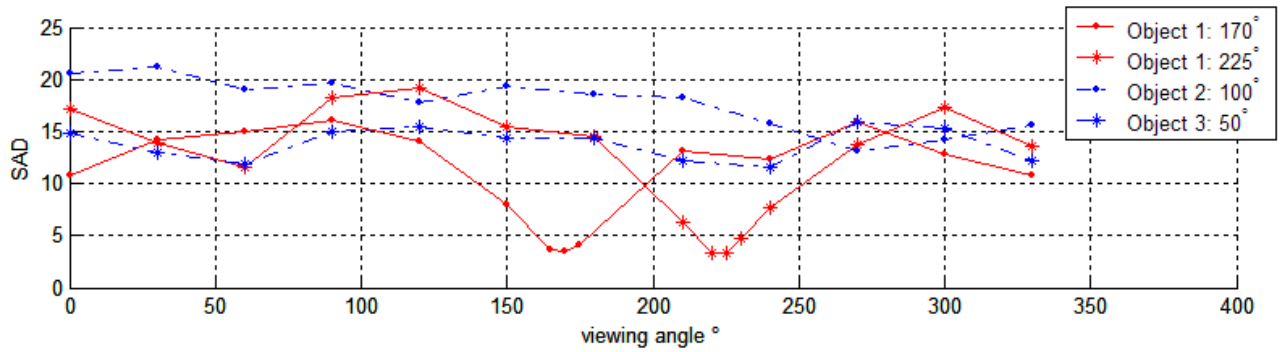


Figure 10: Several ATR results using the view morphing database of object 1. The red lines correspond to two queried images from the same object, which can be correctly identified, and the viewing angles are also estimated (170° and 225° respectively). The blue dot lines correspond to two images from the different objects 2 and 3, whose SADs are not converged to a small value.

between the real images and synthetic images.

Once the view morphing databases are constructed, we can simply identify the input image using the our recognition scheme as mentioned in Section 3. Figure 10 shows that given object 1's view morphing database, the input images can be classified two groups, "correct" and "incorrect". If the input images are from the same object, the SAD value will converge to lower than some threshold (typically 5-6) as the red lines in Figure 10. In addition, the viewing angle of the queried image is the angle corresponding to the minimal value. If the objects do not belong to this object, the SAD values will not converge as shown by the blue dot lines in Figure 10. Therefore, using this scheme, we can easily identify the object only based on the appearance, and even can estimate the viewing angle of the input image. We test all of the images in the testing data set S'' (total 480 images) in this framework, all of the results are correct without the false alarm.

5 Conclusion

In this paper, we present a novel approach to automatically recognize the target based on our multi-view morphing framework. Given a set of reference images of the object, a view morphing database is constructed using the

multi-view morphing algorithm, and a high quality image for an arbitrary novel viewpoint can be generated by extracting the information from the database. Then, using the triple tree search scheme, the input image can be easily identified and the viewing angle can also be estimated.

In the future, we will extend this approach and apply it to video based recognition.

References

[1] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts", IEEE Trans. PAMI, 24(4):509522, 2002.

[2] R. Collins, R. Gross, and J. Shi, "Silhouette-based human identification from body shape and gait", 5th Intl. Conf. on Automatic Face and Gesture Recognition, 2002.

[3] A. Efros, A. Berg, G. Mori, J. Malik, "Recognizing Action at a Distance", International Conference on Computer Vision, 2003.

[4] B. Li, R. Chellappa, Q. Zheng, and S. Der,"Model-Based Temporal Object Verification Using Video", IEEE Trans. Image Processing, 2001.

[5] B. Li, R. Chellappa, Q. Zheng, and S. Der, "Experimental Evaluation of FLIR ATR Algorithms", Computer Vision and Image Understanding, 2001.

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

[7] S. Seitz. *Image-Based Transformation of viewpoint and scene appearance*. Dissertation, computer science, university of Wisconsin, 1997.

[8] P. VanMaasdam, J. Riddle, "A Technique for Constructing an Integated Scene From Multiple Viewing Angles Using a Tactical Ranging Sensor", SPIE Automatic Target Recognition XIII, 2003.

[9] J. Xiao, and M. Shah, "Two-Frame Wide Baseline Matching", International Conference on Computer Vision, 2003.

[10] J. Xiao and M. Shah, "From Images to Video: View Morphing of Three Images", Vision, Modeling, and Visualization (VMV), 2003.

[11] J. Xiao and M. Shah, "Tri-view Morphing", Computer Vision and Image Understanding, 2004.