Face Cyclographs for Recognition*

Guodong Guo†

Department of Computer Science
North Carolina Central University
†*E-mail: gdguo@nccu.edu*

Charles R. Dyer

Computer Sciences Department
University of Wisconsin-Madison
E-mail: dyer@cs.wisc.edu

A new representation of faces, called face cyclographs, is introduced for face recognition that incorporates all views of a rotating face into a single image. The main motivation for this representation comes from recent psychophysical studies that show that humans use continuous image sequences in object recognition. Face cyclographs are created by slicing spatiotemporal face volumes that are constructed automatically based on real-time face detection. This representation is a compact, multiperspective, spatiotemporal description. To use face cyclographs for recognition, a dynamic programming based algorithm is developed. The motion trajectory image of the eye slice is used to analyze the approximate single-axis motion and normalize the face cyclographs. Using normalized face cyclographs can speed up the matching process. Experimental results on more than 100 face videos show that this representation efficiently encodes the continuous views of faces.

Keywords: Face representation; Face cyclograph; Multiperspective; Motion.

1. Introduction

Over the last several years there have been numerous advances in capturing multiperspective images, i.e., combining (parts of) images taken from multiple viewpoints into a single representation that simultaneously encodes appearance from many views. Multiperspective images[1,2] have been shown to be useful for a growing variety of tasks, notably scene visualization (e.g.,

*The support of the National Science Foundation under Grant No. CCF-0434355 is gratefully acknowledged.

panoramic mosaics[3,4]) and stereo reconstruction.[5] Since one fundamental goal of computer vision is object recognition,[6] a question may be asked: are multiperspective images of benefit for object recognition?

Under normal conditions, 3D objects are always seen from multiple viewpoints, either from a continuously moving observer who walks around an object or by turning the object so as to see it from multiple sides. This suggests that a multiperspective representation of objects might be useful.

Recently, psychophysical results have shown that the human brain represents objects as a series of connected views[7,8].[9] In psychophysical experiments by Stone,[7] participants learned sequences which showed 3D shapes rotating in one particular direction. If participants had to recognize the same object rotating in the opposite direction, it took them significantly longer to recognize and the recognition rate decreased. This result cannot be reconciled with traditional view-based representations[10] whose recognition performance does not depend on the order in which images are presented. Instead, it is argued in[7] that the temporal characteristics of the learned sequences such as the order of images are closely intertwined with object representation. These results and others from physiological studies[11] support the hypothesis that humans represent objects as a series of connected views.[9]

The findings from human recognition may have practical guidance for developing better computational object recognition systems. Bülthoff et al.[9] presented a method for face recognition based on psychophysical results[7,8] in which they showed experimentally that the representation of connected views gives much better recognition performance than traditional view-based methods. The main idea of their approach is to process an input sequence frame-by-frame by tracking local image patches to achieve segmentation of the sequence into a series of time-connected "key frames" or views. However, a drawback of the "key frames" representation is that it still needs several view images to characterize the whole sequence.

Can we integrate all continuous views of an object into a *single* image representation? In this paper we propose a new method to incorporate all views of an object, which is the cyclograph of the object[12],[2] a type of multiperspective image. A cyclograph is generated when the object rotates in front of a static camera or the camera rotates around the object. It has a long history in photography, with the first patent related to making cyclographs in 1911.[12] Cyclographs have been used previously for image-based rendering[13] and stereo reconstruction.[5] But, to our knowledge, there is no previous work using cyclographs for object recognition.

The major contributions of this paper are: 1) Propose a new method to incorporate all continuous views of an object for recognition using the cyclograph image of the object. 2) Develop a method to recognize faces using cyclographs based on a dynamic programming technique which can align and match cyclographs simultaneously. 3) Present a method to normalize cyclographs based on motion trajectory image analysis and image warping.

The paper is organized as follows. Section 2 presents the analysis of the *spatiotemporal volume* of continuous views of objects, and the generation of face cyclographs. Section 3 presents two methods for face recognition using face cyclographs. Experimental results are given in Section 4.

2. Viewing Rotating Objects

Our goal is to develop a computational method that encodes all continuous views of faces for face recognition. In psychophysical experiments, the connected views of an object are captured from object rotation in one particular direction[7].[9] Following the psychophysical experiments, we consider the class of single-axis rotations and associated appearances as the basis for capturing the continuous views of faces. The most natural rotations in depth for faces are when an erect person rotates his or her head, resulting in an approximately single-axis rotation about a vertical axis. Many other objects have single-axis rotations as the most "natural" way of looking at them. When we see a novel object we usually do not see random views of the object but in most cases we walk around it or turn the object in our hand.[9]

2.1. *Spatiotemporal Volume*

Suppose that a 3D object rotates about an axis in front of a camera and a sequence of images are captured. Stacking together the sequence of images, a 3-dimensional volume, x-y-t, can be built, which is called a *spatiotemporal volume*.

In psychophysical studies, this 3D volume data is called a *spatiotemporal signature* and there is evidence showing that such signatures are used by humans in object recognition,[14] but no computational representation was presented. We analyze the spatiotemporal volume and generate a computational representation of rotating objects.

The *spatiotemporal volume*, x-y-t, is a stack of x-y images accumulated over time t. Each x-y image contains only appearance but no motion information. On the contrary, the x-t or y-t images contain both spatial and

temporal information. They are called *spatiotemporal images*. The x-t and y-t images can be obtained by slicing the x-y-t volume.

Given a 3D volume, x-y-t, all the x-t (or y-t) slices preserve all the original information without any loss. This is not difficult to see. The x-y slices are captured by the camera, while the x-t or y-t slices are cut from the volume independently. The union of all x-t (or y-t) slices is exactly the original volume. On the other hand, different slices, *i.e.*, x-y, x-t, or y-t, encode different information from the 3D volume.

Although both x-t and y-t slices are *spatiotemporal images*, they contain different information. When the object rotates about an axis that is parallel to the image's y axis, each x-t slice contains information on object points along a horizontal line on the object surface, defining the motion trajectories of these points. On the contrary, each y-t slice contains column-wise appearance of the object surface because of the object rotation about an axis that is parallel to the image's y axis. Thus y-t slices encode the appearance of the object as it rotates 360^o.

When a convex (or nearly convex) object rotates about an axis 360^o, the *spatiotemporal volume* is constructed by stacking the whole sequence of images captured by a static camera. The slice that intersects the rotation axis usually contains the most visible appearance of the object in comparison with other parallel slices. Furthermore, this slice has also least distortion.

2.2. *Spatiotemporal Face Volume*

To represent rotating faces for recognition we need to extract a spatiotemporal sub-volume containing the face region, which we call the *spatiotemporal face volume*. A face detector[15] can be used to automatically detect faces in sequences of face images. The face positions reported by the face detector can then be used to determine a 3D face volume. False alarms from the face detector are removed by using facial skin color information. The eyes, detected with a similar technique as that in the face detector,[15] are used for locating the motion trajectory image of the eye-level slice.

2.3. *Face Cyclographs*

Given a *spatiotemporal face volume* with each coordinate normalized between 0 and 1, we can analyze the 3D face volume via slicing. Based on Section 2.1, one may slice the volume in any way without information loss. However, the y-t slices encode all of the visible appearance of the object for single-axis rotation about a vertical axis. Furthermore, the unique slice

that intersects the rotation axis contains the most visible appearance of the object with minimum distortion among all y-t slices. As a result, we will use this slice for the rotating face representation.

In our face volume, the slice that intersects the rotation axis is approximately the one with $x = 0.5$. This middle slice extracts the middle column of pixels from each frame and concatenates them to create a face-like image, called the "cyclograph of a face," or simply "face cyclograph." One face cyclograph is created for each face video. The size of a face cyclograph image is determined by the video length and the size of the segmented faces, i.e., the width of the face cyclograph is the number of frames in the video, and the height is the height of the segmented faces.

A face cyclograph can also be viewed as being captured by a strip camera,[13] and all parts of the face surface are captured equally well. The face cyclograph representation is compact and multiperspective.

3. Recognition using Face Cyclographs

For face recognition, one face cyclograph is computed for each face video sequence containing one rotating face. Given a testing face sequence, the face cyclograph is computed first and then matched to all face cyclographs in the database. Two algorithms have been developed for matching face cyclographs. The first uses dynamic programming (DP)[16] for alignment and matching of face cyclographs. The monotonicity condition has to be satisfied to use DP and face cyclographs satisfy this by keeping the temporal order of the original face sequences. The second method analyzes the face motion trajectory image and then normalizes face cyclographs to the same size before matching. See[17] for details.

4. Experiments

4.1. *A Dynamic Face Database*

A face video database with horizontal head rotation was captured. Each subject was asked to rotate his or her head from an approximately frontal view to an approximately profile view. 28 individuals, each with 3 to 6 videos, were captured for a total of 102 videos in the database. The number of frames per video varies, ranging from 98 to 290, resulting in a total of 21,018 image frames. Each image is size 720×480.

Each video in our face video database was matched against all other face videos, providing an exhaustive comparison of every pair of face videos. Precision and recall measures were computed with respect to the top n

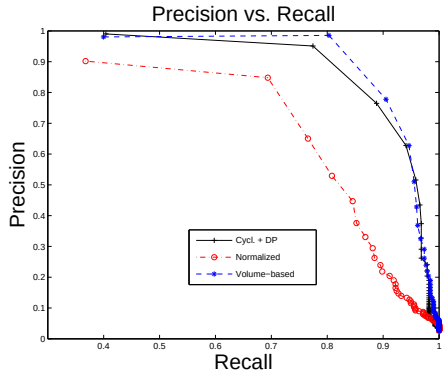

Fig. 1. Average precision versus recall of face cyclographs, face volume-based, and normalized face cyclographs.

matches to evaluate the algorithm's performance.

4.2. *Face Recognition Results*

Face cyclographs were created for all 102 face videos in our database. No faces were missed by this completely automatic process. Given a query face cyclograph, the costs of matching it with all remaining 101 face cyclographs were computed and sorted in ascending order. Then the precision and recall were computed with respect to the top n matches, with $n = 1, 2, \cdots, 101$. Finally, the precision and recall were averaged over all 102 queries and are shown in Fig. 1.

Using the normalized face cyclograph method, the performance was lower than using DP. The reason may be that linear warping introduces artifacts. A non-linear warping method is under consideration.

The face cyclograph algorithms were also compared with a volume-based face recognition method, where the whole face volume was used for matching using the dynamic programming optimization method. As seen in Figure 1, the performance of the face cyclographs methods is very close to the volume-based method in terms of precision and recall. However, using the whole volume has two disadvantages: (1) it requires a large amount of storage, and (2) it is very slow for volume-based matching. In our experiment, the program took more than 24 hours in order to obtain the precision and recall curve (as shown in Figure 1) using the whole volume data as input, while it took just a couple of minutes using the face cyclograph representation.

5. Conclusions

Motivated by recent psychophysical studies, this paper presented a new face representation, called face cyclographs, for face recognition. Temporal characteristics are encoded as part of the representation. This new representation is compact, robust, and simple to compute from a *spatiotemporal face volume*, which itself is automatically constructed from a video sequence. Face recognition is performed using dynamic programming to match face cyclographs or using normalized face cyclographs based on motion trajectory analysis and image warping. We expect the multiperspective representation to improve results of other object recognition tasks as well.

References

1. D. N. Wood, A. Finkelstein, J. F. Hughes, C. E. Thayer and D. H. Salesin, Multiperspective panoramas for cel animation, in *Proc. SIGGRAPH*, 1997.
2. S. M. Seitz and J. Kim, *IEEE Computer Graphics and Applications* **23**, 16(November/December 2003).
3. S. Peleg and J. Herman, Panoramic mosaics by manifold projection, in *Proc. Computer Vision and Pattern Recognition Conf.*, 1997.
4. H. Y. Shum and L. W. He, Rendering with concentric mosaics, in *Proc. SIGGRAPH*, 1999.
5. S. M. Seitz and J. Kim, *Int'l J. of Computer Vision* **48**, 21 (2002).
6. D. Marr, *Vision* (Freeman Publishers, 1982).
7. J. Stone, *Vision Research* **39**, 4032 (1999).
8. G. M. Wallis and H. H. Bülthoff, Effect of temporal association on recognition memory, in *Proc. Natl. Acad. Sci. USA*, 2001.
9. H. H. Bülthoff, C. Wallraven and A. Graf, *Proc. 16th Int. Conf. Pattern Recognition* **3**, 768 (2002).
10. M. J. Tarr and H. H. Bülthoff, *Object recognition in man, monkey, and machine (cognition special issues)* (Cambridge, MIT Press, 1999).
11. Y. Miyashita, *Nature* **335**, 817 (1988).
12. A. Davidhazy, Principles of peripheral photography, in *http://www.rit.edu/ andpph/text-peripheral-basics.html*,
13. P. Rademacher and G. Bishop, Multiple-center-of-projection images, in *Proc. SIGGRAPH*, 1998.
14. J. Stone, *Vision Research* **38**, 947 (1998).
15. P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features, in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.
16. L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition* (Prentice Hall, Englewood Cliffs, 1993).
17. G.-D. Guo, Face, expression, and iris recognition using learning-based approaches, PhD thesis, University of Wisconsin, (Madison, WI, 2006).