Multi-Modal Recognition Using Multiple Views

Gregory Shakhnarovich, Lily Lee & Trevor Darrell

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

Problem: We are interested in combining the recognition results based on gait and face of a person moving in the common field of view of a number of monocular cameras.

Motivation: Person tracking and recognition systems should ideally integrate information from multiple views, and work well even when people are far away. Two key issues that make this challenging are varying appearance due to changing pose, and the relatively low resolution of images taken at a distance. We wish to design a system for real-time multimodal recognition from multiple views that substantially overcomes these two problems.

Broadly speaking, there are several classes of techniques for view-independent face recognition, including modular learning, elastic matching, view-interpolation, and geometric warping. We develop a visual hull approach as an instance of the last category, using multiple views and silhouette inputs.

Previous work: The well-known eignenfaces paradigm was extended to recognize a set of different poses using an eigenspace for each view [4]. Rather than use replicated classifiers for distinct views, several authors have investigated elastic matching (e.g. [6]) or view interpolation methods [1]. A view morphing technique was developed in [5], but it was not appllied for recognition.

Generalizing the notion of elastic matching, recognition based on principle components analysis of shape and texture distributions has been shown to be able to model and recognize a range of object poses[2]. When a model has been constructed fast optimization of shape and texture coefficients is possible. However, all these methods have generally presumed either knowledge of face pose and/or an accurate, dense depth or correspondence field during model training. This can be difficult to acquire in practice.

For tracking, several authors have used planar as well as affine, cylindrical and ellipsoidal models to warp views and bring images into a canonical view. However, these simple static models are often inaccurate and difficult to align with dynamically changing observations.

Approach: To deal with the varying appearance due to changing pose, we adopt a view-normalization approach, and use an approximate shape model to render images for recognition at canonical poses. These images are sent to externally provided recognition modules which assume view-dependent input. For distant observations view-normalization must not presume accurate 3-D models are available; our system is designed for environments where relatively coarse-disparity stereo range images or segmented monocular views are provided.

Low-resolution information makes recognition using any single modality less accurate. By combining cues into a multimodal scheme, we can obtain increased performance. A typical drawback of multimodal approaches is that they presume different types of imagery as input. Face recognition usually works best with front-parallel images of the face, whereas gait recognition often requires side-view sequences of people walking. It can be difficult in practice to simultaneously acquire those views when the person is moving along a variable path. Our proposed view-normalization method performs this automatically, generating appropriately placed virtual views for each modality.

We use a shape model based on an efficient technique for computing image-based visual hulls ([3]), and recognition algorithms separately developed for view-dependent recognition. In our system a small number of static cameras observe a workspace and generate segmented views of a person; these are used to construct a 3-D visual hull model. Canonical virtual camera positions are estimated, and rendered images from those viewpoints are passed to the recognition methods. Figure 1 gives an example of the an input images observed by one of the cameras (top row), and the virtual views rendered for gait (middle) and face (bottom) recognition.

Figure 1: Input on one of the cameras (top), and synthethic views rendered as an input to gait (middle) and face recognition algorithms.

Impact: A view-independent multimodal recognition method allows to identify people with high confidence under conditions of varying pose. Possible application of such ability include, but are not limited to, identification and tracking of users in intelligent environment, surveilance, and safety monitoring.

Future Work: Currently the implementation uses monocular silhouettes based on color segmentation with static backgrounds, but could be extended to accommodate more sophisticated segmentation algorithms. Our system works within the strict intersection of the field of view of all cameras, but we expect this to be relaxed as a more general visual hull algorithm is developed. Finally, our confidence integration method is clearly primitive in present form, and should be extended to an explicit probabilistic framework. The classification scheme also may include additional modalities, besides gait and face appearance. Another promising direction of research is to use the multiview framework for articulated body tracking, evntually leading to detection and analysis of activities.

Research Support: This project is funded with the generous support of the Darpa HumanID program and MIT Project Oxygen.

References:

[1] D. Beymer and T. Poggio. Face recognition from one example view. In *Proccedings of the International Conference on Computer Vision*, pages 500–507, 1995.

[2] G. J. Edwards, T. F. Cootes, and Christopher J. Taylor. Face recognition using active appearance models. In *ECCV (2)*, pages 581–595. Springer, 1998.

[3] Wojciech Matusik, Chris Buehler, and Leonard McMillan. Polyhedral visual hulls for real-time rendering. to appear in *Proceedings of EGWR-2001*, 2001.

[4] A. Pentland, B. Moghaddam, T. Starner, O.Oliyide, and M. Turk. View-based and modular eigenspaces for face recognition. Technical Report 245, MIT Media Lab Vismod, 1993.

[5] S. M. Seitz and C. R. Dyer. Photorealistic scene reconstruction by voxel coloring. In *Proceedings Computer Vision and Pattern Recognition (CVPR'97)*, pages 1067–1073, 1997.

[6] L. Wiskott, J.M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7):775-779, 1997.