Learning Silhouette Features for Control of Human Motion

Liu Ren
Carnegie Mellon University

Gregory Shakhnarovich
MIT

Jessica Hodgins
Carnegie Mellon University

Hanspeter Pfister
MERL

Paul Viola
Microsoft Research

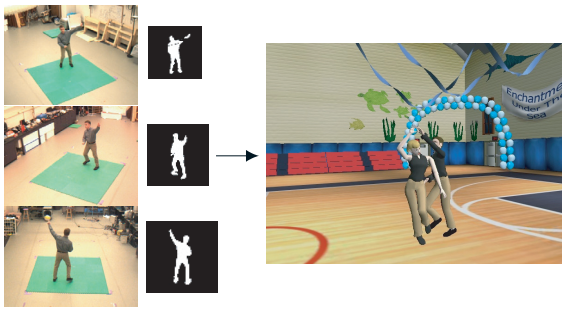

Figure 1: A dancing user drives two animated characters.

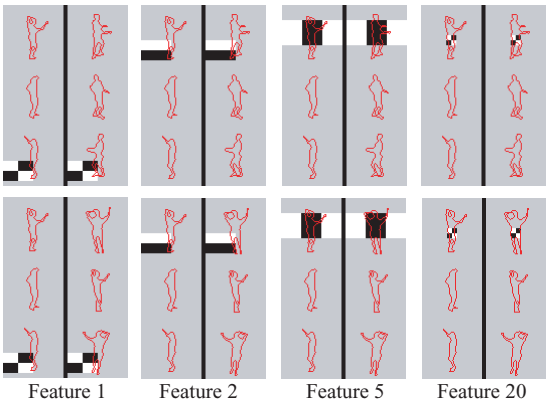

Figure 2: Sample local features selected by AdaBoost for yaw estimation. Each local feature can be located at any position with different scales and aspect ratios. The top row shows positive training examples (pairs of silhouettes) with similar yaw angles. Negative examples are shown in the second row. The silhouettes are shown with contour lines for better visualization.

People, even without training, can move in expressive ways to play charades, pantomime a winning touchdown, or act out a story for a child. The goal of our research is to provide untrained users with a vision-based interface for interactively controlling complex human motions. The insight behind our system is that the domain knowledge from a pre-recorded motion database allows plausible movement to be inferred from the incomplete information obtained from simple vision processing. The experimental scenario we use in this research is a single user animating two people swing dancing. We present a low cost camera-based system (one PC and three video cameras) that allows an untrained user to "drive" the dance of the couple interactively (Figure 1). The user can change the high-level structure of the dance, introducing a different sequence of steps, rotations, or movements. The final sequence retains the quality of previously captured dance yet reflects the user's higher-level goals and spontaneity. Our approach yields results for more complex behaviors and longer sequences than have been demonstrated in previous silhouette-based systems.

We rely on discriminative learning to allow the vision-based interface to accurately map silhouette inputs to complex human motions such as swing dancing. The vision-based interface relies on silhouettes extracted from three video streams to match the animator's body against a database of poses and movements. This matching process uses a set of discriminative local features that are computed on the silhouette images [Viola and Jones 2001]. These features are computationally efficient and therefore suited to real-time applications. The best local features for estimating yaw and body configuration are selected from a very broad set of possible features based on a variant of the AdaBoost algorithm [Schapire and Singer 1999] (Figure 2). Synthetic data are used for the feature selection.

The determination of the yaw orientation incorporates a fast search method called locality sensitive hashing. The determination of the body configuration (joint angles, root position, and orientation) relies on the temporal coherence in a domain-specific database of human motion. The database is pre-processed into an augmented motion graph that extends the motion graph [Lee et al. 2002] to handle scaling in velocity and two interacting characters. The search for the pose of the animated character is performed locally in the augmented motion graph in the region around the frame currently being animated. A look-ahead buffer provides sufficient information for the match and introduces a delay of less than one second.

Our experimental results include individual users dancing and the resulting motions of a couple performing East Coast Swing. For testing, we capture video and motion capture data simultaneously and compare the motion computed by our system with the ground-truth motion capture data. We also compare the performance of the learned local features with that of a commonly used global feature, Hu moments, to demonstrate the effectiveness of discriminative local features for mapping 2D silhouettes to 3D poses.

We envision that a system such as this one might provide the infrastructure for teleconferencing applications, an interactive game for a dancing child, an animated karaoke performance for an adult, or with a sufficiently broad database, an animation authoring interface for a naive user. Alternative approaches to creating performance-based animation systems include vision-based tracking algorithms, small marker sets in concert with inverse kinematics, real-time motion capture systems, and sensors such as accelerometers or a foot pad. Our system has potential advantages over these approaches because it requires no instrumentation of the user (other than reasonably tight-fitting clothing and fixed lighting), produces high quality motion given an appropriate database, and requires only off-the-shelf cameras rather than special-purpose sensors or cameras. For more information, see graphics.cs.cmu.edu/projects/swing.

References

LEE, J., CHAI, J., REITSMA, P., HODGINS, J., AND POLLARD, N. 2002. Interactive control of avatars animated with human motion data. In *ACM Transactions on Graphics*, vol. 21, 491–500.

SCHAPIRE, R. E., AND SINGER, Y. 1999. Improved boosting algorithms using confidence-rated predictions. In *Machine Learning*, 297–336.

VIOLA, P., AND JONES, M. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.