Video Based Animation Techniques for Human Motion

Chris Bregler, Stanford University

Most image based rendering techniques are applied to rigid domains: Static environment maps, indoor scenes, or architectural scenes. Explicit geometric structures are combined with image data. Texture mapping and view morphing are simple examples. We can generate new images from a collection of recorded images. Simple geometry dictates coarse transformations of fine grained image texture. New views of a scene can be generated in blending between the transformed example textures. This is a trade-off between explicit structure (collection of views and geometric model) and implicit example data (the image texture).

Such trade-offs are applied to other domains as well. The most successful speech production systems (text-to-speech, concatenative speech) follow a similar philosophy. A collection of annotated example sounds are used to create new sounds. A sentence is build from phonemes (explicit structure). To blend the phonemes together, the sound examples are pitch and time warped (implicit data). We will show how this extends to video data and human motion animation.

Structure vs Data for Animation:

So far most graphical animation techniques do not exploit such trade-offs between explicit structure and implicit data. Many facial and body animations are generated by 3D volumetric models and physical simulations. Some facial animation systems texture map images onto the geometric model, or morph between a few example images. The appearance and motion become increasingly realistic. We will survey some of these systems.

In contrast to physical simulations, motion capture based animation techniques become increasingly popular. An actor performs the desired motion, and devices record body joint configurations or facial configurations. This data is mapped onto graphical computer models. Motion editing techniques allow to modify the motion data and create new motions. This has similarities to image morphing techniques. Instead of warping image texture, spatio-temporal configurations are warped.

Some systems allow to blend between different motion-capture data sets of different actions. New animations are assembled using existing examples.

Video Based Animation of People:

In order to create animations, that have natural motion **AND** have photo-realistic appearance, we need to combine motion-capture and image based (or video based) techniques. The goal is to build video based representations of annotated example motions.

Unlike standard motion capture techniques that are based on markers or other devices, we need to annotate body and facial configurations directly in unconstrained video. In static scenes the user could supply annotations by hand, but for video sequences, automatic techniques are crucial (10 min of video has 18,000 images, no-one has the budged, patience, and consistency to do this by hand). We will survey several visual tracking and annotation techniques that are tailored for full body movements and facial

movements. We demonstrate these visual annotation techniques on lab recordings of people walking and talking. We also demonstrate how to process historic footage. Examples are the famous Edweard Muybridge Plates from over 100 years ago of walking people, and stock-footage of John F. Kennedy giving a public speech.

To build libraries of example motions, we also need techniques that annotate coarse motion categories automatically. Again, this has to be done automatically. For example a 10 minute video of someone talking could be transformed into a video-based library of more then 2,000 phonetic lip motions (phonemes or visemes).

Once they are annotated, we can re-animate the data or create new data. We will present work-in-progress of photo-realistic animations of kinematic chain models, and we will cover in more detail work presented at last years SIGGRAPH conference on photo-realistic animation of talking heads: Video Rewrite.

More technical details of such techniques can be found in following papers:

- Video Rewrite: C. Bregler, M. Covell, M. Slaney

  Video Rewrite uses existing footage to create automatically new video of a person mouthing words that she did not speak in the original footage. This technique is useful in movie dubbing, for example, where the movie sequence can be modified to sync the actors' lip motions to the new soundtrack.
  Video Rewrite automatically labels the phonemes in the training data and in the new audio track. Video Rewrite reorders the mouth images in the training footage to match the phoneme sequence of the new audio track. When particular phonemes are unavailable in the training footage, Video Rewrite selects the closest approximations. The resulting sequence of mouth images is stitched into the background footage. This stitching process automatically corrects for differences in head position and orientation between the mouth images and the background footage.
  Video Rewrite uses computer-vision techniques to track points on the speaker's mouth in the training footage, and morphing techniques to combine these mouth gestures into the final video sequence. The new video combines the dynamics of the original actor's articulations with the mannerisms and setting dictated by the background footage.
  Video Rewrite is the first facial-animation system to automate all the labeling and assembly tasks required to resync existing footage to a new soundtrack.

- Video Motion Capture: C. Bregler, J. Malik

  This paper demonstrates a new vision based motion capture technique that is able to recover high degree-of-freedom articulated human body configurations in complex video sequences. It does not require any markers, body suits, or other devices attached to the subject. The only input needed is a video recording of the person whose motion is to be captured. For visual tracking we introduce the use of a novel mathematical technique, the product of exponential maps and twist motions, and its integration into a differential motion estimation. This results in solving simple linear systems, and enables us to recover robustly the kinematic degrees-of-freedom in noise and complex self occluded configurations. We demonstrate this on several image sequences of people doing articulated full body movements, and visualize the results in re-animating an artificial 3D human model. We are also able to recover and re-animate the famous movements of Eadweard

Muybridge's motion studies from the last century. To the best of our knowledge, this is the first computer vision based system that is able to process such challenging footage and recover complex motions with such high accuracy.

Preliminary slides in PDF

References

- [Beier92] T. Beier, S. Neely. Feature-based image metamorphosis. SIGGRAPH 92, 26(2):35-42,1992.
- [Ezzat98] T. Ezzat, T. Poggio. MikeTalk: A Talking Facial Display Based on Morphing Visemes. Proc. Computer Animation Conference, Philadelphia, Penssylvania, 1998. (CD-version)
- [Guenter98] B. Guenter, C. Grimm, D. Wood, H.Malvar, F.Pighin. Making Faces. SIGGRAPH 98, 55-66. (web-pointer)
- [Litwinowicz94] P. Litwinowicz, L. Williams. Animating images with drawings. SIGGRAPH 94, Orlando, FL, pp. 409-412,1994
- [Moulines90] E. Moulines, P. Emerard, D. Larreur, J.L. Le Saint Milon, L. Le Faucheur, F. Marty, F. Charpentier, C. Sorin. A real-time French text-to-speech system generating high-quality synthetic speech. Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Albuquerque, MN, pp. 309-312, 1990.
- [Pighin98] F. Pighin, J. Hecker, D.Lischinski, R.Szeliski, D.H.Salesin. Synthesizing Realistic Facial Expressions from Photographs.
- [Scott94] K.C. Scott, D.S. Kagels, S.H. Watson, H. Rom, J.R. Wright, M. Lee, K.J. Hussey. Synthesis of speaker facial movement to match selected speech sequences. Proc. Australian Conf. Speech Science and Technology, Perth Australia, pp. 620-625, 1994. (CD-version)
- [Water95] K. Waters, T.Levergood. DECface: A System for Synthetic Face Applications. J. Multimedia Tools and Applications, 1(4):349-366, 1995.(web-pointer)
- [Williams90] L. Williams. Performance-Driven Facial Animation. SIGGRAPH 90, 24(4):235-242, 1990.