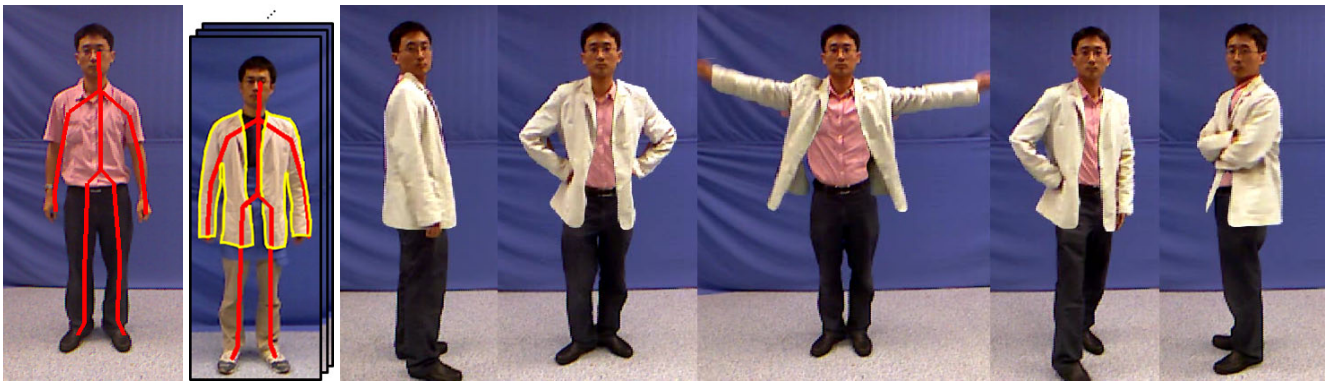
Image-based Clothes Animation for Virtual Fitting

Zhenglong Zhou[1] Bo Shu[1] Shaojie Zhuo[1] Xiaoming Deng[1,2] Ping Tan[1] Stephen Lin[3]
[1]National University of Singapore [2] Institute of Software, Chinese Academy of Sciences [3]Microsoft Research Asia



(a) (b) (c)

Figure 1: *Virtual clothes fitting. (a) Sample frame in the input video, where the user's skeletal pose is recorded by a motion capture device. (b) Segmented garments from a database indexed from pre-recorded video clips according to the user's skeletal pose. (c) Selected output video frames in which the virtual garment is displayed over the user's figure for virtual clothes fitting.*

Abstract

We propose an image-based approach for virtual clothes fitting, in which a user moves freely in front of a virtual mirror (i.e., video screen) that displays the user wearing a superimposed virtual garment. The motions and deformations of the virtual garment are synthesized by tracking the user's skeleton and utilizing corresponding video clips of the actual garment worn by a model. Our system overcomes challenges in generating a convincing garment animation. With our technique, we developed a real-time system using a Microsoft Kinect camera that demonstrates effective clothes fitting results with a variety of garment types.

Links: DL PDF

1 Introduction

Virtual clothes fitting applications allow users to see how they appear in chosen clothes without physically putting them on. Typically, a camera captures the user standing in front of a screen that shows a real-time visualization of him/herself wearing the virtual garment. Some existing commercial systems such as Swivel [SWIVEL 2011] overlay a still image of the garment on the user's figure. To provide clothes animation as the user moves, other systems such as the Fitnect [Fitnect 2011] resort to 3D garment modeling, rendering and animation, each of which is a challenging problem itself. For example, the complexity of wedding dresses with fine decorations makes them difficult to model and animate. In addition, fabrics such as velvet and satin have special reflectance properties that are challenging to reproduce. Because of these issues, it is hard to generate photorealistic feedback in real-time with a 3D modeling approach, and this may degrade the user experience.

We advocate an image-based approach for clothes animation. Under fixed lighting and viewing conditions, we address the kinematics of clothes appearance including its effects on reflectance and shape deformation. From a dataset of captured video clips with a garment in different configurations and undergoing various motions, we synthesize arbitrary animations of the garment by rearranging the temporal order of video frames through an optimization framework. However, a major challenge is that only a small subset of the possible garment configurations/motions can practically be captured on video from a clothing model, whose natural motions typically differ from that of the user. As a result, it is often difficult to rearrange video frames to generate garment animations that accurately follow the user's movements. We present a skeleton based warping technique and optical flow based frame interpolation to address this problem to provide the user with a visually natural clothes fitting experience. This approach is implemented using a Microsoft Kinect camera to measure character poses. Our results demonstrate the effectiveness of our system in generating real-time photorealistic feedback to users.

2 Related Works

Video textures was introduced by Schödl et al. [2000] to model stochastic motions of volumes such as fire and water. Novel sequences are generated by rearranging and assembling existing video footage. This idea was generalized in 'video sprites' [Schödl and Essa 2002] to create controlled animal animations. Face animations have also been created from existing video footage based on similar ideas, such as in 'video rewrite' [Bregler et al. 1997] and 'video puppetry' [Brand 1999].

Improved frame similarity measures were designed to ap-

ply this data-driven approach for full body character animation [Celly and Zordan 2004; Flagg et al. 2009; Starck et al. 2005; Xu et al. 2011]. Though these methods have the potential to be used for clothes animation, as clothes are part of characters, they are not immediately applicable. As explained in [Flagg et al. 2009], the silhouette based method in [Celly and Zordan 2004] cannot accurately capture pose similarity. [Flagg et al. 2009] relied instead on optical markers to facilitate pose estimation and frame warping, but this causes distracting artifacts in clothes and cannot be applied to loose-fitting garments. [Starck et al. 2005; Xu et al. 2011] recovered 3D character models, but 3D garment reconstruction is difficult, especially for loose-fitting clothes [Stoll et al. 2010]. Our work instead focuses on clothes animation and does not rely on optical markers or fragile 3D reconstruction.

Virtual clothes fitting is a relatively recent application. [Hilsmann and Eisert 2009] virtually changed the texture pattern of a T-shirt while maintaining correct clothes deformation and shading. Our work is most similar to [Hauswiesner et al. 2011b], which also uses an image-based approach with a multi-camera setup to transfer clothes. There are some key differences between it and our method. For example, Hauswiesner et al. [2011b] relied on 3D models – image based visual hulls – for visualization, while our method is completely image based. Furthermore, Hauswiesner et al. [2011b] only selected some distinct frames to minimize the garment database size, while we keep all the video frames and furthermore generate novel motions for smooth and accurate clothes animation. All these differences are important for successfully generating vivid animations of loose-fitting clothes (see Figure 1). By contrast, Hauswiesner et al. [2011b] only demonstrated tight-fitting garments with limited realism. The authors later extended their work in [Hauswiesner et al. 2011a] to drive a 3D character model with a Kinect camera for online shopping applications.

Virtual fitting applications are also found outside of the academic literature. Existing systems can be roughly divided into 2D and 3D systems according to their data representation. Swivel [SWIVEL 2011] deforms a still image to fit a skeleton to allow pose changes. 3D virtual fitting systems were demonstrated by [Fitnect 2011], where 3D garment models are created and animated according to the user's motion. In general, 2D systems do not provide realistic garment animation, while 3D systems face challenges including modeling, rendering and animation. We chose an image-based approach to achieve realistic animations without going through the 3D modeling and rendering pipeline.

3 Data Preparation

In building the video clip database for a given garment, we use a blue screen background to facilitate garment segmentation. The clothes models were asked to perform some common clothes fitting actions, as well as move freely to provide additional garment configurations and motions. We record approximately 5,000 video frames for each garment, together with the corresponding skeletal poses. The animated garment is captured by a color video camera calibrated and synchronized with the motion capture device. We use the 'Roto Brush' in Adobe After Effects to interactively segment the garment in these database video frames, which takes about 30 minutes for each garment.

In principle, the body shape of the model should match that of the user for the superimposed garment to fit well. Since this is generally not the case, image warping [Jain et al. 2010] could be applied to the database videos according to an estimate of the user's body shape [Weiss et al. 2011]. For simplicity, we assume here that body shape of the user and model are about the same, and leave body shape modification for future work.

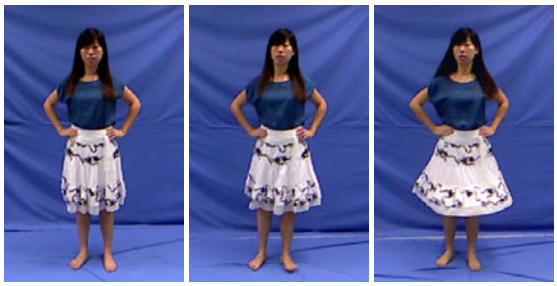


Figure 2: *Frames where models have the same pose but different motions.*

4 Garment Transfer

In generating the output video, our system uses the user's skeletal pose to query the database for garment frames captured with a similar pose. For this, we present an algorithm that optimizes pose match while promoting consistency between neighboring frames. We further address issues with limited and noisy data.

We first present the basic framework of our system. The video database is indexed by a *pose vector*, a representation of the skeletal pose by a concatenation of its 3D joint positions. For each input video frame, we query the database to find frames with a similar skeletal pose. To minimize sudden changes in the resulting animation, we also favor selection of frames that form a contiguous sequence in a database clip. These two criteria are joined in the following objective function:

$$E = \sum_i D(l_i, i) + \lambda \sum_i S(l_{i-1}, l_i), \tag{1}$$

where l_i is the index of the selected database frame for the i-th frame in the input video. $D(l_i, i)$ is the Euclidean distance between the pose vectors of l_i and i. In the second term, $S(l_{i-1}, l_i)$ is set to zero when l_{i-1} and l_i are neighboring frames in a database clip; otherwise it is set to one. To account for different motion speeds between the model and user, we consider l_{i-1} and l_i to be neighboring frames if they are separated by fewer than four frames in the database video (i.e., we allow up to a $\times 4$ speedup). λ is a fixed parameter (100 in our implementation).

This objective function E is defined on a one dimensional chain, so we can apply the dynamic programming (DP) algorithm to quickly obtain its global optimum. Since DP requires later frames to determine the result at the current frame, we buffer several (M=10) frames and apply DP to the frames within the interval $[i-1, i+M]$ to obtain the result at the i-th frame. For real-time performance, it is generally infeasible to exhaustively check all database frames in optimizing the objective function E. So for each input frame, we first apply approximate nearest neighbor (ANN) search [Arya et al. 1998] to find a set of N candidate frames (N=75 in our implementation) with the most similar pose vectors. DP then solves for the most appropriate one from among these candidates.

Once an appropriate garment frame is found, the segmented garment is superimposed as virtual clothing on the user in the input video frame. The garment is positioned by matching the mean position of the hip and shoulder joints of the user's and model's skeletons. Its scale is set according to relative shoulder width.

Motion-aware frame query Especially for loose-fitting clothes, the clothes deformation depends not only on skeletal pose but also

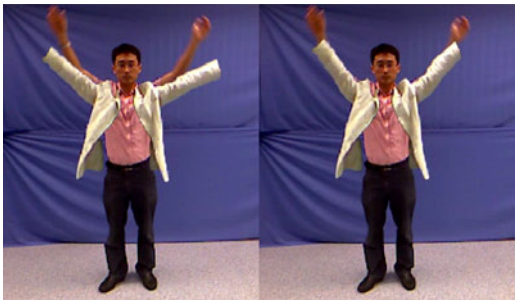


Figure 3: *The left and right are results without and with warping.*

on motion. As shown in Figure 2, when the user turns around slowly or quickly, the skeletal pose is the same. However, the motion of the skirt differs significantly, flying up only when the user turns quickly. Hence, we include motion information in the query to better model clothes deformations. We simply concatenate pose vectors in neighboring frames (11 frames in our implementation), since motion can be derived from consecutive poses. We refer to this concatenated vector as a *motion vector*, and use it in place of the pose vector in the baseline method. In addition, the term $S(\cdot, \cdot)$ is set to the difference between the motion vectors of l_{i-1} and l_i when l_{i-1} and l_i are not neighboring frames in a database clip.

Skeleton based warping Since database video clips provide only a sparse sampling of poses in the pose space, the garment obtained by optimization often has a configuration that does not exactly match the pose of the user. We address this problem by performing a non-rigid deformation of the garment to improve alignment. For runtime efficiency, we do not consider full body shape warping as done in [Jain et al. 2010]. Instead, we use the skeleton joints on the four limbs as control points to warp the garment image to better fit the user's pose. The algorithm described in [Schaefer et al. 2006] is used for warping. Figure 3 shows an example of how this skeleton based warping leads to an improved fit.

Frame interpolation and alignment Though the dynamic programming optimization favors smooth motion between neighboring video frames, jittering may appear at the connection between two contiguous sequences selected from the database. Suppose $A_1, A_2, \cdots A_n$ are the n database frames selected by the dynamic programming from a contiguous sequence, and $B_1, B_2, \cdots B_m$ similarly form the following contiguous sequence. For a smoother transition between A_n and B_1, we compute optical flow [Brox et al. 2004] between them and linearly interpolate corresponding points to generate the resulting frame B_1'. We also interpolate B_1' and B_2 to create B_2' to replace B_2. This interpolation is performed for K frames, where $K = 5$ in our implementation.

5 Experiments

In our implemented system, a Kinect camera is set in front of the character. This camera is used to obtain garment images and record character poses. Our algorithm is run on an eight-core PC with a 2.8G CPU and 12G RAM, which yields performance at 25 frames per second. Our motion vector representation is compact enough to load the motion vectors of all database frames into RAM. To expedite loading of database garment images, we pre-load the first 10,000 frames into RAM during system initialization. At run time, whenever a garment image outside of RAM is required, we load its neighboring frames to replace the 2000 oldest frames in RAM.

Figure 1 shows some results on a loose-fitting suit. (a) is a sample frame from the input video with the skeleton overlaid on it. (b) represents a set of segmented database frames with their associated skeletons. Several output frames with different poses are exhibited in (c). Please view the supplementary video for animation results. Our image-based approach convincingly models clothes animation and reflectance, which is difficult to achieve through 3D modeling and rendering. Compared with the existing image-based clothes transfer work [Hauswiesner et al. 2011b][1], our method demonstrates smoother clothes animation on more challenging loose garments. More results are provided in Figure 4.

6 Conclusion and Discussion

We proposed an image-based method for virtual clothes fitting. We first build a segmented garment database indexed by skeleton pose and motion. At run time, we search for suitable database frames and overlay them on the user's figure. The result is an animated virtual clothes fitting system, which produces vivid animations and realistic appearance at real-time rates. This method may also be applied to dress up virtual characters in computer generated animations.

In our experiments, we observed several limitations of our system, which point to directions for further improving our method. First, visible discontinuities sometimes exist in the output clothes animation, due mainly to imprecise pose recovery from the Kinect cameras and insufficient sampling of the pose space. Use of a professional motion capture device may significantly improve results, especially when the character turns around. There is much current research [Girshick et al. 2011] aimed at improving pose estimation from Kinect cameras, and our work can greatly benefit from advancements in this area. Second, our current system only provides a fixed viewpoint visualization of the garment. If viewpoint change is desired, we could set an array of cameras surrounding the model to capture a multi-view garment database and apply the same database query technique. Lastly, our skeleton based warping could generate artifacts when limbs overlap with the body region. However, we could adopt the automatic segmentation proposed in [Flagg et al. 2009] to separate limbs and body.

7 Acknowledgement

This project is partially supported by the NSFC project No. 61005039 and Singapore grant R-263-000-698-305.

References

ARYA, S., MOUNT, D. M., NETANYAHU, N. S., SILVERMAN, R., AND WU, A. Y. 1998. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *J. ACM*, 891–923.

BRAND, M. 1999. Voice puppetry. In *Proc. of SIGGRAPH*, 21–28.

BREGLER, C., COVELL, M., AND SLANEY, M. 1997. Video rewrite: driving visual speech with audio. In *Proc. of SIGGRAPH*, 353–360.

BROX, T., BRUHN, A., PAPENBERG, N., AND WEICKERT, J. 2004. High accuracy optical flow estimation based on a theory for warping. In *In Proc. of ECCV*, Springer, 25–36.

CELLY, B., AND ZORDAN, V. B. 2004. Animated people textures. In *Proc. of International Conference on Computer Animation and Social Agents*.

FITNECT. 2011. http://www.fitnect.com/.

[1]Please refer to their video demos at http://www.icg.tugraz.at/project/narkissos/publications.

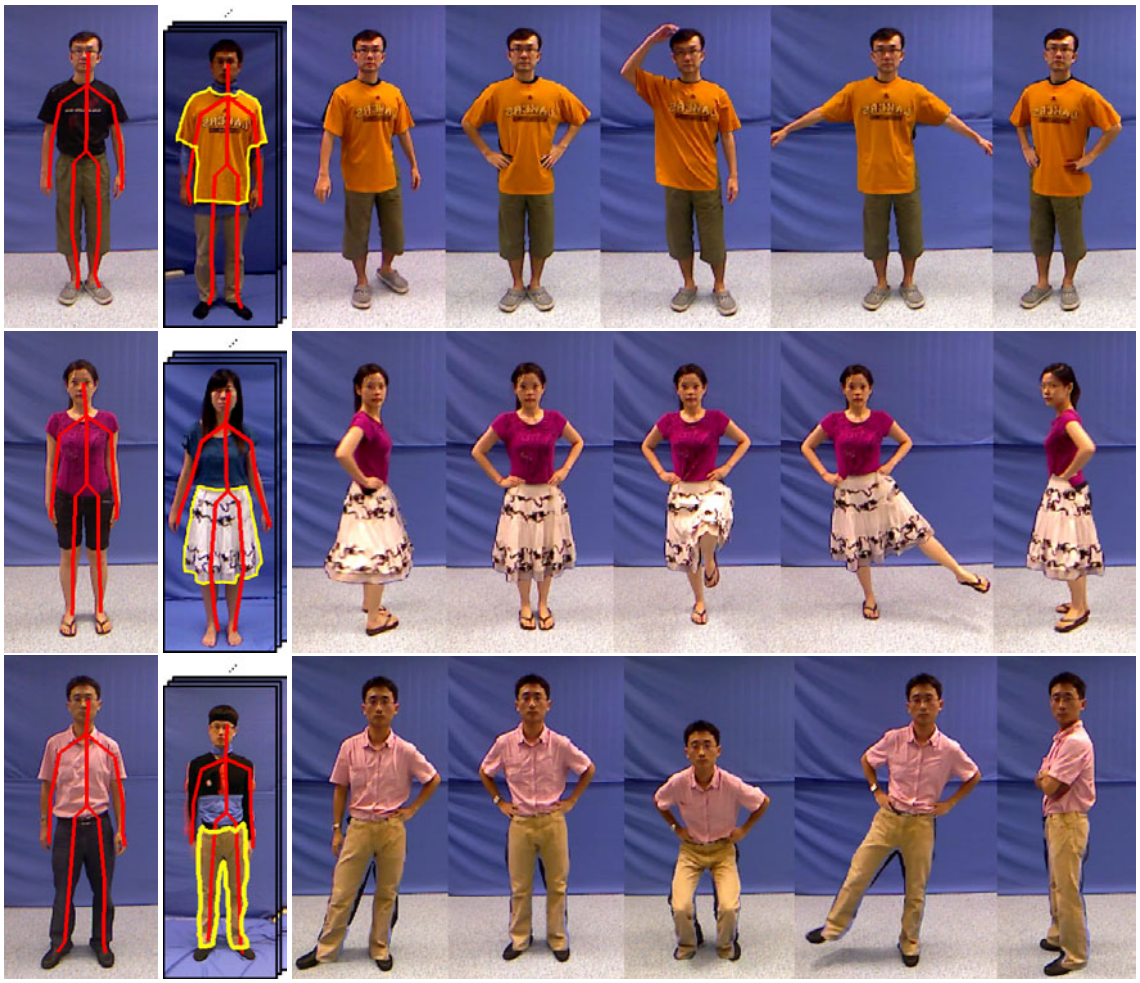


Figure 4: *From left to right are an input video frame, the corresponding segmented garment from the database, and sample output frames.*

FLAGG, M., NAKAZAWA, A., ZHANG, Q., KANG, S. B., RYU, Y. K., ESSA, I., AND REHG, J. M. 2009. Human video textures. In *Proc. of symposium on Interactive 3D graphics and games (I3D)*, 199–206.

GIRSHICK, R., SHOTTON, J., KOHLI, P., CRIMINISI, A., AND FITZGIBBON, A. 2011. Efficient regression of general-activity human poses from depth images. In *Proc. of ICCV*.

HAUSWIESNER, S., STRAKA, M., AND REITMAYR, G. 2011. Free viewpoint virtual try-on with commodity depth cameras. In *Proc. of International Conference on Virtual Reality Continuum and Its Applications in Industry (VRCAI)*.

HAUSWIESNER, S., STRAKA, M., AND REITMAYR, G. 2011. Image-based clothes transfer. In *Proc. of International Symposium on Mixed and Augmented Reality (ISMAR)*.

HILSMANN, A., AND EISERT, P. 2009. Tracking and retexturing cloth for real-time virtual clothing applications. In *Proc. of Mirage*, 94–105.

JAIN, A., THORMÄHLEN, T., SEIDEL, H.-P., AND THEOBALT, C. 2010. Moviereshape: tracking and reshaping of humans in videos. *ACM Trans. Graph. 29*.

SCHAEFER, S., MCPHAIL, T., AND WARREN, J. 2006. Image deformation using moving least squares. *ACM Trans. Graph.*.

SCHÖDL, A., AND ESSA, I. A. 2002. Controlled animation of video sprites. In *Proc. of Eurographics symposium on Computer animation*, 121–127.

SCHÖDL, A., SZELISKI, R., SALESIN, D. H., AND ESSA, I. 2000. Video textures. In *Proc. of SIGGRAPH*, 489–498.

STARCK, J., MILLER, G., AND HILTON, A. 2005. Video-based character animation. In *Proc. of Eurographics symposium on Computer animation*, 49–58.

STOLL, C., GALL, J., DE AGUIAR, E., THRUN, S., AND THEOBALT, C. 2010. Video-based reconstruction of animatable human characters. *ACM Trans. Graph. 29*.

SWIVEL. 2011. http://www.facecake.com/.

WEISS, A., HIRSHBERG, D., AND BLACK, M. 2011. Home 3d body scans from noisy image and range data. In *Proc. ICCV*.

XU, F., LIU, Y., STOLL, C., TOMPKIN, J., BHARAJ, G., DAI, Q., SEIDEL, H.-P., KAUTZ, J., AND THEOBALT, C. 2011. Video-based characters: creating new human performances from a multi-view video database. *ACM Trans. Graph. 30*.