Low Cost Video Animation of People using a RGBD Sensor

Cathrine J. Thomsen, Thomas B. Moeslund and Troels H. P. Jensen
Analysis of People Lab, Aalborg University, Rendsburggade 14, 9000, Aalborg, Denmark

Keywords: Kinect Sensor, Video Animation, Performance Capture, Segmentation, Optical Flow.

Abstract: This paper is an investigation in a low cost solution for performing video animation using a Kinect v2 for Windows, where skeleton, depth and colour data are acquired for three different characters. Segmentation of colour and depth frames were based on establishing the range of a person in the depth frame using the skeleton information, and then train a plane of the floor and exclude points close to it. Transitioning between motions were based on minimizing the Euclidean distance between all feasible transitioning frames, where a source and target frame would be found. Intermediate frames were made to create seamless transitions, where new poses were found by moving pixels in the direction of the optical flow between the transitioning frames. The realism of the proposed animation was verified through a user study to have a higher rate of preference and perceived realism compared to no animation and animation using alpha blending.

1 INTRODUCTION

Animating people by using marker-based motion capture has been widely used and is perhaps most known for the Gollum character in *The Lord of the Rings*, where the actor wears a tight suit with reflective markers. Tracking each of these markers in a multiple-view studio, when the actor performs different motions, are then used for controlling the joints of an animated character performing the same motions. Not only does the capture using the marker-based approach require a significant setup time, but it also takes the actor or actress out of the natural environments and lacks surface details, such as the dynamics of the hair and clothing.

Instead, marker-less animation has been introduced, where different motion sequences are captured in a multiple-view studio as in (Xu et al., 2011) and (De Aguiar et al., 2008) using HD cameras. Using a multiple-view stereo approach, 3D meshes are then reconstructed independently for each frame, which means that both shape and appearance of the captured person are preserved.

Recent work within reconstruction and modelling of people using multiple cameras can result in a realistic yet expensive video animation. The focus of this paper is to investigate a low-cost solution using a consumer video and depth cameras, in this case a Kinect v2 for Windows, capturing sequences to make a realistic animation of a person, which will build on previous work within 4D animation. The main contributions of this work are the segmentation and animation methods in this low-cost solution.

The next section presents the related work within this topic and the rest of the paper is divided into five sections. The first section includes how the datasets are captured, then the segmentation and the animation methods are described. The paper is finalized by a user study and a conclusion.

2 RELATED WORK

Capturing various motions of a person, thus having a library of different motions, new animations can thereby be created by combining and transitioning between related motions in the library. Using a motion graph (Kovar et al., 2002), a user has the possibility to control the different movements that a character should perform. This motion graph is an animation synthesis that controls feasible transitioning points between chosen motions, where all states in the graph consist of small clips in a video library of that particular motion.

In order to establish the best transition between two motions, a similarity measure between the feasible transitioning frames is made. According to a study of different similarity measures in (Huang et al., 2010), the shape histogram is the similarity measure that proved to give the best performance between dif-

Thomsen C., Moeslund T. and Jensen T.
Low Cost Video Animation of People using a RGBD Sensor.
DOI: 10.5220/0006132902440249
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 244-249
ISBN: 978-989-758-224-0

ferent people and motions, also used in (Budd et al., 2013). The shape histogram subdivides a spherical coordinate system into radial and angular bins, where each bin represents the mesh in that specific area. The similarity measure between two meshes compares all rotations of the mesh around the centroid for which the L2 distance is minimum, thus being rotation invariant.

By having a reference mesh, it is possible to compare how well the reconstruction of the surface and texture of a character is (Casas et al., 2013). But whether the animations are made in a realistic way depends on how it is perceived. This means, it is hard to establish a measure for the realism of the animated videos without testing how people perceive them. Hereby, the approach in (Casas et al., 2014), where a user study was conducted to test the realism of their videos, would be an appropriate way for assessing an animation procedure.

3 DATA CAPTURE

A dataset consisting of four characters performing various motions is used in this work, which are shown in figure 1 and figure 2, where figure 1 is used for evaluation and the others are used for testing. The data was captured with 30 fps using the Kinect v2 for Windows sensor, which includes a colour data stream with a resolution of 1920x1080, a depth data stream with a resolution of 512x324, and finally the skeletal tracking from the Windows SDK v2.0 consisting of 25 joints per person.

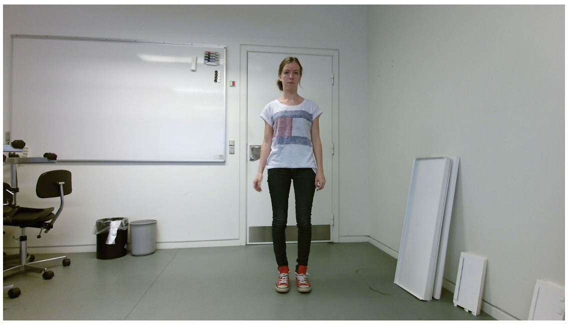

Figure 1: Colour frame 100 of the character used for evaluation waving alternately left and right hand.

Having one or multiple video clips for one particular motion, a motion graph as in (Starck and Hilton, 2007) is constructed, thus having a database of each motion for each character. This motion graph is then used for controlling the transition between each motion within each character by keeping track of the current state and the possible transitions. Since the captured datasets in this work consists of a full video of a person performing various motions, the motion sequences are divided and labelled manually by determining the end and start frame in each of the videos. The video is divided in such a manner that each motion sequence do not overlap.

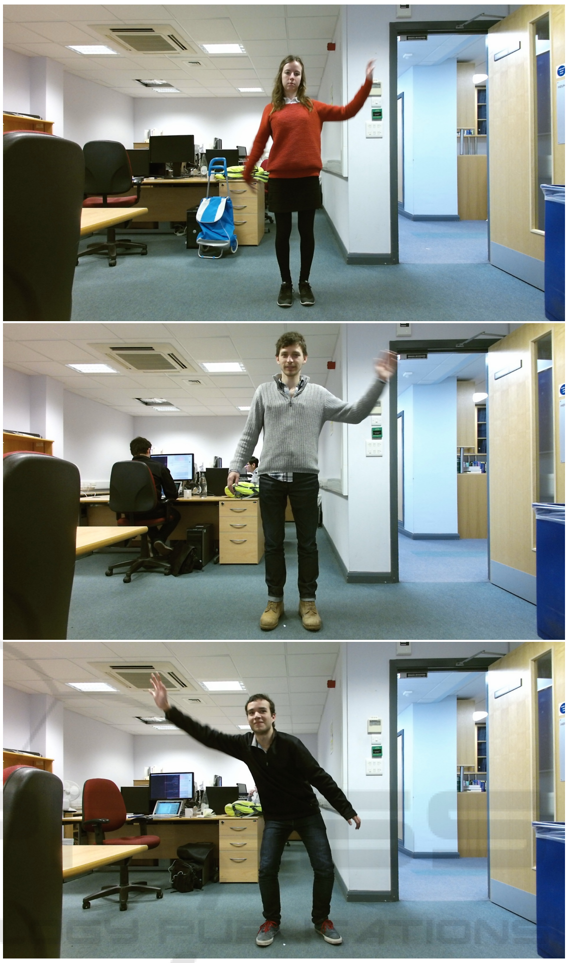

Figure 2: Colour frame 100 of the characters used for testing. From the top is character 1 performing dance motions, character 2 performing stretches and lastly is character 3 performing goalkeeper motions.

4 SEGMENTATION

Since the background is of no interest when creating an animation of a person, a segmentation must be performed. The segmentation can be done using the colour or depth frames. If the colour frame is used, illumination changes between the frames as well as background cluttering must be taken into account. Another solution is to use the depth frames instead, as in (Fechteler et al., 2014), which avoids the cluttering

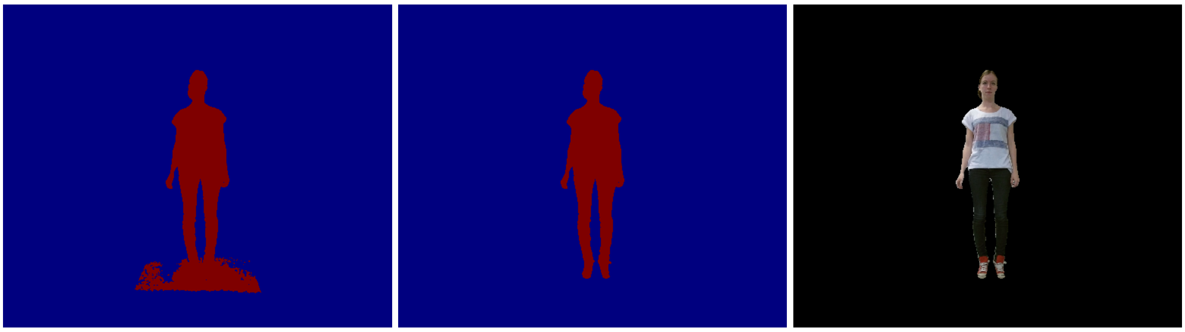

Figure 3: Segmented depth frame using the information from the skeletal joints on the left, and the resulting segmented depth and colour frame.

and contains less illumination issues. Since the camera is static, a depth segmentation could be done by modelling the background of an empty scene. But in this case, it does not provide a good enough segmentation, therefore another approach must be used.

4.1 Plane Segmentation

With inspiration from the grund plane segmentation in (Møgelmose et al., 2015), a similar depth-based approach for the segmentation is proposed. Instead of establishing a bounding box around the person by e.g. HOG-detectors on RGB images, a simpler and faster approach is proposed by using the depth image only, since the skeleton of the person is available. The range of the person is therefore established by using the minimum and maximum positions of the joints from the skeleton in each direction in the depth image giving the output shown in figure 3.

Having the range of the person, a plane from the floor is created by the span of the minimum and maximum value in the x-direction and the minimum value in the z-direction, as illustrated in figure 4.

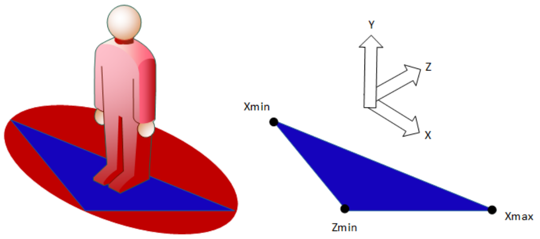

Figure 4: Illustration of the extraction of the floor plane.

The distance from the plane, α, with a normal vector, (a,b,c), to a point, (x_0,y_0,z_0) is evaluated from equation 1, where the points close to the plane are excluded so only the person remains.

$$dist_{\alpha,point} = \frac{|a(x-x_0)+b(y-y_0)+c(z-z0)|}{\sqrt{a^2+b^2+c^2}} \quad (1)$$

Since small pixel changes can occur between each depth frame, an average of planes is used instead of calculating a new plane for each frame. Experimental results showed that using 20 frames was sufficient to construct an average plane. Additional noise left in the frames were removed by preserving only the biggest BLOB using a 4-connectivity connected component analysis.

The final segmentation is shown in figure 3. Here a boolean bitwise AND operation is applied to the segmented output frame and the original depth frame. This depth frame is then further used for mapping the colour frame onto the depth frame. Thereby, the resulting segmented colour frame has the same resolution as the depth frame of 512x424.

5 TRANSITIONING BETWEEN MOTIONS

In order to create an animated video, where a segmented person is shifting between different motions, good transitions between the motions needs to be determined. The three modalities captured with the Kinect v2 for Windows; skeleton, depth and colour, are used for creating a similarity measure. This similarity measure is used for determining good transitions between different frames in the video where shape and appearance have a good match.

For each modality, the similarity is calculated as the Euclidean distance between two feature vectors. In terms of the colour and depth data, the feature vectors correspons to the unravelled pixels in a frame, whereas the skeleton's feature vector corresponds to the location of each of the 25 joints.

After individual normalization of each similarity measure by dividing all values with the maximum value, they are combined into one similarity measure, where the weight for each measure are α, β and γ:

$$S = \alpha S_{skel} + \beta S_{depth} + \gamma S_{colour} \quad \alpha, \beta, \gamma \in [0,1] \quad (2)$$

When determining a possible transition between two motions in a motion graph, a source and a target motion are established. Minimizing the similarity measure, S, between the two motions will then find the two frames with the best transition, as illustrated in figure 5.

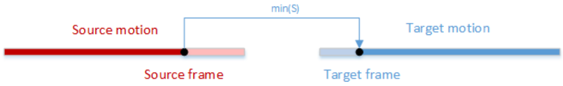

Figure 5: Illustration of searching for the best transition between two motions is where the similarity measure is minimized over all frames.

If multiple motions are added to a video, the target motion then becomes the new source motion, and going through the motion graph and the database of motions, a new set of one or more possible target motions are available. Finding the best transition to the next motion is done, yet again, as illustrated in figure 6, where all frames in the dark red bar are a part of the resulting video.

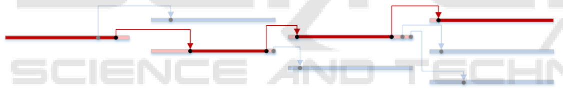

Figure 6: Illustration of searching for the best transitions through multiple motions and the resulting video sequence marked with dark red.

5.1 Smoothing Transitions

Since the source and target frames can have a difference in shape and appearance when transitioning between these frames, it will create a jump in the resulting video, which is unwanted. An example is given in figure 7.

Smoothing the transitions will therefore be needed, which can be done by creating new intermediate frames between the source and target frames. Creating intermediate frames using alpha blending will result in ghosting (Casas et al., 2014) when the shape is not similar, as shown in the top of figure 8, which is unwanted.

Instead, the sparse optical flow between the source and target frame is established by using the implementation of Farnebäck's optical flow (Farnebäck, 2003) in OpenCV 2.4.10. The displacement for each individual pixel between the source and target frame are then gradually moved along this displacement vector to perform the animation, which is shown in figure 9.

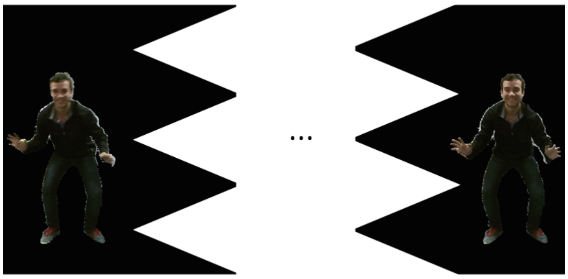

Figure 7: A source and target frame for the best transition between two motions where $\alpha = \beta = 1$ and $\gamma = 0$.

Due to rounding errors, when doing a zeroth-order interpolation of the flow vector, holes appear in the resulting frame. It also happens, when a pixel does not have a destination of a flow vector from another pixel. This makes sense when e.g. moving the arm, where the previous position of the arm should be empty in the new frame. But this can also happen inside the body, which must be handled. One approach to avoid this is to do backward mapping using the inverse transformation. Although, if a pixel does not have a destination from the optical flow, it has no transformation thus no inverse transformation.

Thereby, the empty pixels that need to be filled after a forward mapping are found by evaluating the surrounding pixels in the output frame via a median filter. The resulting animation compared to the alpha blending is shown in figure 8.

6 USER STUDY

In order to evaluate the perceived realism, when transitioning between motions, a user study is performed. 10 different settings are each compared for 3 different characters.

These settings, described in table 1, include different weightings of the three modalities for the similarity measure, and it also includes transitions with no animation and with animation based either on alpha blending or optical flow. The first, second and third column is the value for the α, β and γ respectively in equation 2, and the last corresponds to the number of intermediate animated frames inserted between a transition, where ”$*$” is animation using only alpha blending.

21 non-experts took part in a web-based survey, where the test subject was presented with two randomized video pairs, i.e. 45 video pairs for each character. Each pair was to be rated from 1 (top preference)

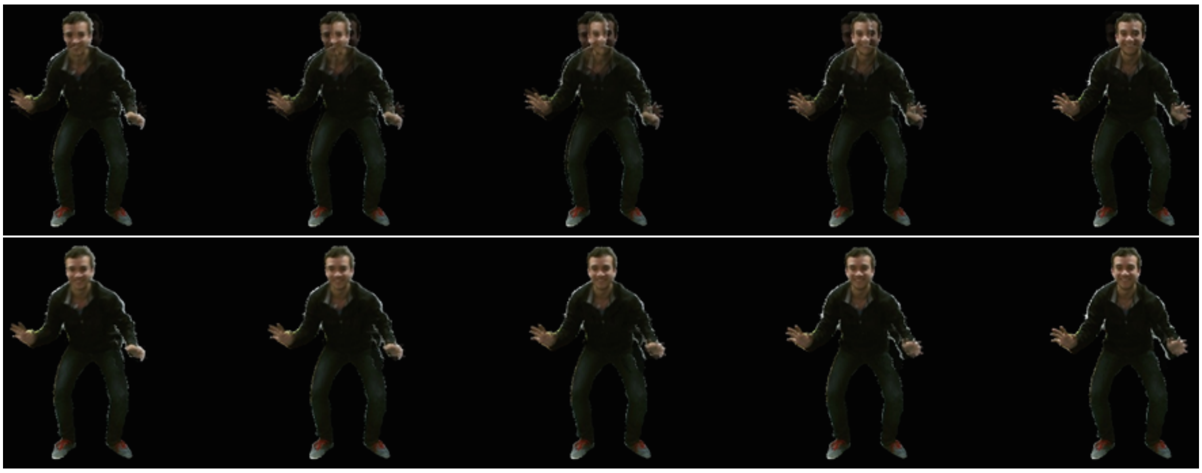

Figure 8: Animation between the source and target frame from figure 7, the top is based on alpha blending and the bottom is based on the proposed method.

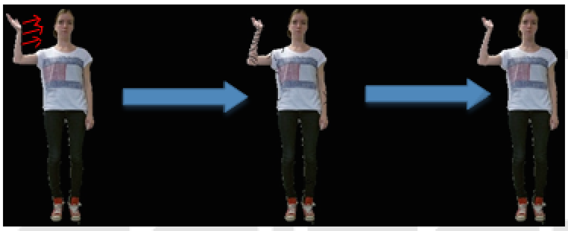

Figure 9: Illustration of how an intermediate frame is made based on the proposed method by using optical flow.

Table 1: The parameter settings for each video for each character. *Animated frames based on alpha blending.

α (skeleton)	β (depth)	γ (colour)	frames
1	-	-	-
1	-	-	5
1	1	-	-
1	1	-	5
1	1	-	5*
1	-	1	5
1	1	1	5
-	1	-	5
-	-	1	5
-	1	1	5

to 5 (bottom preference), where 3 was chosen if there was no preference.

The results obtained for each character are shown in figure 10, where each setting from table 1 on the x-axis is labelled with four digits.

Interestingly, the result show that the transitions with animation based on optical flow using only the colour data in the similarity measure, received a very low score, and for the character 2, it even got the lowest score. The results also show that using the skeleton information only or adding it to the colour and depth information gave a higher perceived realism than other settings.

The results show that the animation based on alpha blending was significantly rated as one of the lowest for all three characters, and this is followed by the transitioning with no animation, which also received a low score. As an example, using the skeleton and depth information only with no animation, in figure 7, shows a difference in shape, and therefore creates a jump in the output video.

7 CONCLUSIONS

Four datasets of people performing various motions were captured by acquiring the skeleton, depth and colour data from a Kinect v2 for Windows sensor. The motions sequences in each capture were manually divided and inserted in a motion graph to assure only feasible transitions between motions would happen.

Segmenting the colour and depth frames by learning a background model proved to be a poor segmentation, thus another approach was suggested. Using the skeleton information to establish the range of the body in the depth frame gave a segmentation where the person and a part of the floor remains. The floor was then removed by averaging a plane over the first 20 frames of a captured video and excluding points close to the plane in the rest of the video.

The best transition between a source and target motion was found by a similarity measure by minimizing the Euclidean distance of a combination of skeleton, depth and colour data between all possible frames. In order to create seamless transitions between motions, intermediate frames were needed. The sug-

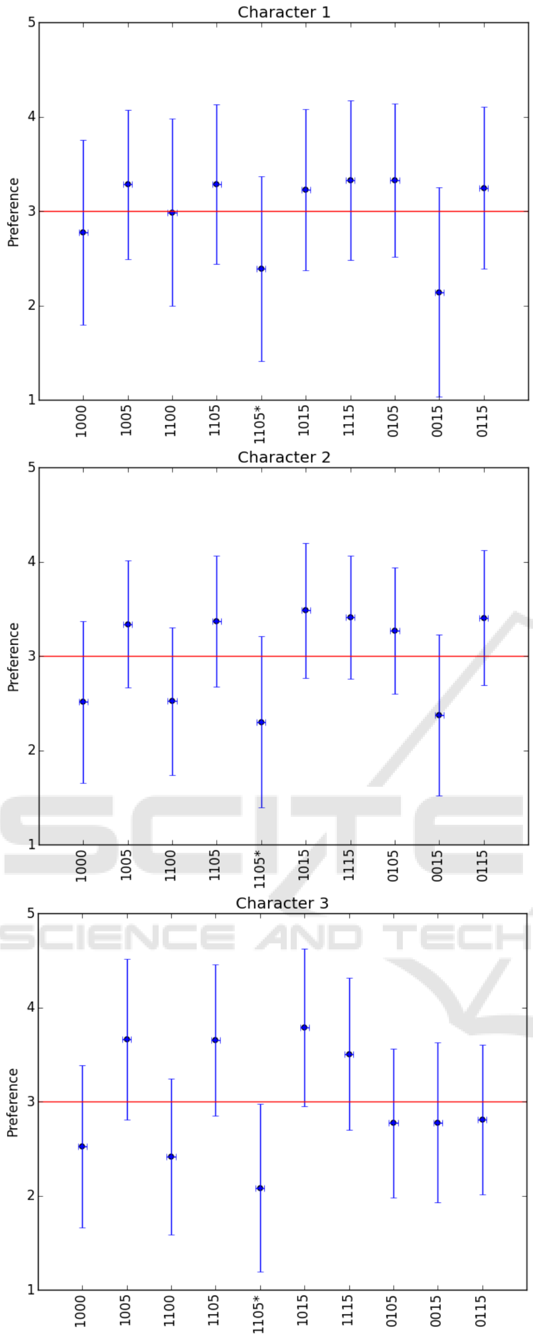

Figure 10: Results for each character obtained from the user study showing the mean and variance for each setting.

gested approach was to create new poses by estimating the dense optical flow between the transitioning frames in each direction and then move the pixels step wise towards the flow and blend the frames from each direction.

The suggested animation based on optical flow was compared with animations using a direct alpha blending between the transitioning frames and not performing any animation. This was evaluated through a web based user study between three different characters. The results showed consistency over all characters, where the suggested animation with different similarity measure settings had a higher rate of preference and perceived realism than using animations based on alpha blending and using no animation.

Creating an adaptive number of intermediate frames, which should vary according to the the pose similarity and speed of motion before and after transitioning, could be a possible area for future investigation.

REFERENCES

Budd, C., Huang, P., Klaudiny, M., and Hilton, A. (2013). Global non-rigid alignment of surface sequences. *International Journal of Computer Vision*, 102(1-3):256–270.

Casas, D., Tejera, M., Guillemaut, J.-Y., and Hilton, A. (2013). Interactive animation of 4d performance capture. *IEEE transactions on visualization and computer graphics*, 19(5):762–773.

Casas, D., Volino, M., Collomosse, J., and Hilton, A. (2014). 4d video textures for interactive character appearance. In *Computer Graphics Forum*, volume 33, pages 371–380. Wiley Online Library.

De Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.-P., and Thrun, S. (2008). Performance capture from sparse multi-view video. In *ACM Transactions on Graphics (TOG)*, volume 27, page 98. ACM.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer.

Fechteler, P., Paier, W., and Eisert, P. (2014). Articulated 3d model tracking with on-the-fly texturing. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 3998–4002. IEEE.

Huang, P., Hilton, A., and Starck, J. (2010). Shape similarity for 3d video sequences of people. *International Journal of Computer Vision*, 89(2-3):362–381.

Kovar, L., Gleicher, M., and Pighin, F. (2002). Motion graphs. In *ACM transactions on graphics (TOG)*, volume 21, pages 473–482. ACM.

Møgelmose, A., Bahnsen, C., and Moeslund, T. B. (2015). Comparison of multi-shot models for short-term re-identification of people using rgb-d sensors. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications 2015*.

Starck, J. and Hilton, A. (2007). Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31.

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 Transactions on Graphics (TOG)*, 30(4):32.