MULTIVIEW FOREGROUND SEGMENTATION USING 3D PROBABILISTIC MODEL

Jaime Gallego, Montse Pardàs

Technical University of Catalonia (UPC)
{jgallego,montse}@gps.tsc.upc.edu

ABSTRACT

We propose a complete multi-view foreground segmentation and 3D reconstruction system that defines a 3-dimensional probabilistic model to model the foreground object in the 3 spatial dimensions, thus gathering the information from all the camera views. This 3D model is projected to each one of the views in order to perform the 2D segmentation with the foreground information shared by all the cameras. Then, for each one of the views, a MAP-MRF classification framework is applied between the projected region-based foreground model, the pixel-wise background model and the region-based shadow model defined for each view. The resultant masks are used to compute the next 3-dimensional reconstruction. This system achieves correct results by reducing the false positive and false negative errors in sequences where some camera sensors can present camouflage situations between foreground and background. Moreover, the use of the 3D model opens possibilities to use it for objects recognition or human activity understanding.

Index Terms— 3D reconstruction, Multi-view foreground segmentation, 3D probabilistic model, SCGMM.

1. INTRODUCTION

3D reconstruction from multiple calibrated planar images is a major challenge in the image processing area in order to obtain a realistic volumetric representation of the objects and people under study. When working with multi-view sequences, it is possible to establish a collaboration between views in order to increase the robustness of the overall system (foreground segmentation + 3D reconstruction). The proposals presented in the literature, usually have an independent processing for each one of the views, and try to improve the final results by combining the probabilities of each view or by using the back projection of the resultant 3D reconstructions [7]. In this paper, we propose a complete integration of the multi-view smart-room segmentation and 3D reconstruction. We propose to define a 3-dimensional modeling of the foreground object under analysis in order to centralize the probabilistic information of the object, for all the views, in the 3-dimensional space, thus giving robustness to the process. This model will be used to achieve the objects' segmentation in each view, preserving the robustness of the model in those views where foreground and background present high similarity and also, it can be exploit to achieve 3D information of the object's movements.

In this system, we define a probabilistic 3D model of the foreground object, where the 3D spatial-color Gaussian Mixture Model (3D SCGMM) is defined to model the probabilistic information of the foreground object to segment in the $v = (RGB, XY, Z)$ domains. This model will be used as a non-rigid characterization of the object. Therefore, in order to correctly define this model, the 3-dimensional reconstruction of the object under analysis and the texture that this object presents in the multi-view sequence are necessary.

1.1. Previous work

In the recent years, there have been special interest in monitoring the human activities and movements in order to obtain a semantic information of the scene. Hence, approaches based on rigid human body models have been proposed in the literature to deal with this analysis. Human motion capture has been extensively studied, [13, 12, 15] give and in-depth survey of the literature. In [5], the multi-layer framework is proposed by means of particle-based optimization related to estimate the pose from silhouette and color data. The approaches in [1, 11, 14] require training data to learn either restrictive motion models or a mapping from image features to the 3D pose. In [16] the authors propose a rigid human body model that comprises a kinematic skeleton and an attached body approximation modeled as a Sum of Gaussians where 58 joints work together to model a detailed spine and clavicles. In [8] shape and motion retrieval are detected by means of EM framework to simultaneously update a set of volumetric voxel occupancy probabilities and retrieve a best estimate of the dense 3D motion field from the last consecutive frame set. In 3-dimensional reconstruction, there have been also great interest in improving the volumetric reconstruction by combining information among views. [3] proposed a *SfS* using Dempster-Shafer theory, which takes into account the positional relationships between camera pairs and voxels to determine the degree in which a voxel belongs to a foreground object. [4] proposed the Space occupancy grids where each pixel is considered as an occupancy sensor, and the visual hull computation is formulated as a problem of fusion of sensors with Bayesian networks, while [10] worked with the Shape from Inconsistent Silhouette by combining the probabilities of each one of the pixels.

1.2. Proposed method

In this paper, we propose a multi-view foreground segmentation method for smart-room scenarios that uses a 3-dimensional probabilistic models to model the object to segment. The work flow of the proposed system is as follows:

Create 3D model: Once all the cameras of the multi-view system have detected and segmented the object under analysis, the foreground 3D SCGMM can be created with the 3D reconstruction obtained from the 2D silhouettes. Although any SfS technique can be used to perform the volumetric reconstruction, we utilize a conservative Visual Hull reconstruction with tolerance $\tau = 1$ in order to reduce the possible misses without increasing too much the false positive detections. Moreover, the voxels of this volume are colorized with the object colors in order to obtain a realistic volume reconstruction, by obtaining the average color that the pixels belong-

ing to the voxel's projection present in each view. The voxels spatial and color information will be used to initialize the foreground 3D SCGMM by means of the EM algorithm [2]. Next frames of the sequence will utilize the 3D model in the segmentation process.

Foreground segmentation: Foreground segmentation is computed by means of the system proposed in [6], thus combining in a Bayesian MRF-MAP framework pixel-wise background model with SCGMM and SCGM foreground and shadow models respectively.

3-dimensional volumetric reconstruction: As in the 3D model creation, conservative Visual Hull reconstruction with tolerance $\tau = 1$ is used in order to obtain the 3D reconstruction of the foreground object that will result the output of the system.

Spatial updating of the 3D model: The 3D object reconstruction will be used to update the 3D foreground model in order to adapt it to the movements that the foreground object performs at each frame. If the model is correctly initialized in the color and spatial domains, only a spatial updating will be necessary to achieve a correct characterization of the object since, unlike the 2D SCGMM, the 3D reconstruction does not present regions occluded to the camera.

Projection of the 3D SCGMM to 2D views: The final step of this work-flow consists in projecting the 3D SCGMM to each one of the views, in order to use the 3D model in the 2D foreground segmentation. Therefore, for each camera sensor, the 2D foreground model will be composed by the projection of the 3D Gaussians that model voxels which present direct visibility from the camera sensor.

The remainder of the paper is organized as follows: Section 2 describes the 3D foreground model. Section 3 explains the projection of the 3D SCGMM to the 2D views. Finally, some results and conclusions are presented in Section 4 and Section 5 respectively.

2. 3D FOREGROUND MODEL

In order to utilize the data redundancy that appear among views, we propose to characterize the foreground object by defining a 3D spatial probabilistic model. This model will gather all the information of the object under analysis, thus increasing the robustness of the multi-view segmentation process.

Since the foreground objects that appear in scene are constantly moving and changing along the sequence, we propose the 3D SCGMM at region based level to model the spatial (XYZ) and color (RGB) domains of the 3D object volume

Therefore, at each time t of the multi-view sequence, our objective is to obtain an updated model parameter set:
$\theta \equiv \{\hat{\omega}, \hat{\mu}, \hat{\Sigma}\} \equiv \{(\omega_1, \mu_1, \Sigma_1) \dots (\omega_k, \mu_k, \Sigma_k) \dots (\omega_{K_{3D}}, \mu_{K_{3D}}, \Sigma_{K_{3D}})\}$, that maximizes the foreground volume (V_t) data likelihood:

$$\theta_{V_t} = \arg\max_{\theta_{V_t}} \prod_{v_i \in V_t} \left[P(v_i|\theta_{V_t})\right], \tag{1}$$

where ω is the Gaussians weight component, μ is the mean, Σ denotes the variance, $v_i \in \mathbb{R}^6$ is the input feature vector for voxel i in the $v = (RGB, XYZ)$ domain and $P(v_i|\theta_{V_t})$ is the likelihood of voxel i formulated as follows:

$$P(v_i|\theta_{V_t}) = \sum_{k=1}^{K_{3D}} \omega_k G_{\text{fg}}(v_i, \mu_k, \Sigma_k), \tag{2}$$

where K_{3D} is the total number of Gaussians that belong to the foreground 3D SCGMM model and $G_{\text{fg}}(v_i, \mu_k, \Sigma_k)$ denotes the pdf of the k-th Gaussian formulated as:

$$G_{\text{fg}}(v_i, \mu_k, \Sigma_k) = \frac{1}{(2\pi)^3 |\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{(v_i - \mu_k)^T \Sigma_k^{-1} (v_i - \mu_k)}{2}\right], \tag{3}$$

where $\mu_k \in \mathbb{R}^6$ is the mean of the 3D Gaussian and $\Sigma_k \in \mathbb{R}^{6\times 6}$ denotes its Covariance matrix.

The 3D SCGMM presents a non-flexible 3D modeling, thanks to the free movement that the 3D Gaussians present, thus adapting well to the real shape of the object without having any movement restrictions.

2.1. Initialization

An initial segmentation of the foreground object in each view is required in order to achieve its first 3D reconstruction. In order to achieve it, we use the planar foreground segmentation system proposed in [6] in each one of the views. Once the foreground object has been initialized and segmented in all the views, we use conservative Visual Hull reconstruction with tolerance $\tau = 1$, in order to achieve the voxelized 3D volume. This volume is colorized assigning to each voxel belonging to the surface of the volume, the color of the 2D pixels correspondent to the voxel projection.

Given this initial colored volumetric reconstruction, the foreground model parameter estimation can be reached via Bayes' development with the EM algorithm ([2]) in the (RGB, XYZ) domains. For this aim, we use only the surface voxels of the volume, since they are the only ones with useful information for the multi-view segmentation analysis, and thus, this will speed up the process.

We estimate how many Gaussians are needed for correctly modeling the object analogously to the proposal presented in [6], i.e. by analyzing the color histogram for this purpose.

After the initialization of the 3D SCGMM, next frames of the sequence will be processed by projecting this 3D foreground model to each one of the views. Hence, in frame t, we will use the projection of the model obtained from $t-1$, to carry out the 2D planar detection in each view. These planar foreground masks will make possible to achieve the 3D SfS reconstruction for frame t, which will be used, in turn, to update the 3D SCGMM before analyzing the next frame of the sequence.

2.2. Updating

The foreground objects perform some displacements and rotations along the scene that makes necessary the model updating at each frame. Since the probabilistic model works in the 3D (XYZ) domain, and the color of the object is correctly modeled from the initialization in the overall volume, only spatial updating is the necessary along the frames. We propose to update the components of the 3D Gaussian Mixture in the spatial domain, for frame t, in a two-steps updating, by using the 3D volumetric reconstruction obtained in the previous step.

2.2.1. Spatial Domain Updating

We use the color and spatial information of the voxels classified as foreground to update only the spatial components of the Gaussian Mixtures. Similarly to the initialization step, we will work with the surface voxels of the 3D volume. Hence, we assign each voxel to the Gaussian k that maximizes:

$$P(k|v_i, \theta_{V_t}) = \frac{P(v_i|\theta_{V_t}, k)}{\sum_k P(v_i|\theta_{V_t}, k)} = \frac{P(v_i|\theta_{V_t}, k)}{P(v_i|\theta_{V_t})}, \tag{4}$$

where $P(v_i|\theta_{V_t})$ is the likelihood of the foreground model for the voxel i (defined in Equation 2), and $P(v_i|\text{fg}, k)$ is the likelihood given by the Gaussian k. Once each voxel has been assigned to a Gaussian, the spatial mean and covariance matrix of each one are

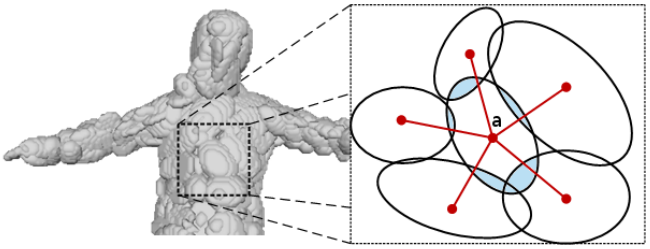


Fig. 1. Example of neighborhood and connectivity between Gaussians that belong to the 3D foreground model. In blue color, the overlapped volume regions between the ellipsoid a under analysis, and the neighbor Gaussians.

updated with the spatial mean and variances of the surface voxels that each one is modeling.

Regularization of the Gaussians displacements:

Once each Gaussian has been spatially updated, we regularize the displacements that each one suffers in the 3D space by using the information obtained from the neighbor Gaussians, thus achieving a more homogeneous spatial evolution of the 3D SCGMM. Hence, given the foreground parameter set $\theta_{V_{t-1}}$ before the spatial updating, and the parameter set after the updating: θ_{V_t}, we calculate the spatial displacements $d_{s=x,y,z} = (d_x, d_y, d_z)$ of the Gaussian k by computing: $d_{s,k} = (\mu_{s,k,t} - \mu_{s,k,t-1})$.

We define this neighborhood according to the connectivity that each one presents in the surface of the volume with respect to the rest of the Gaussians. If we establish the 3D spatial representation of each Gaussian, as an ellipsoid whose axis (ϵ) are defined by the three eigenvalues of its spatial covariance matrix ($\lambda_1, \lambda_2, \lambda_3$) as: $\epsilon_i = 2\sqrt{\lambda_i}$, then two Gaussians will be connected if both present an overlapped region of their spatial ellipsoids (formulated in Cartesian coordinates as: $\frac{(x-\mu_X)^2}{\epsilon_1} + \frac{(y-\mu_Y)^2}{\epsilon_2} + \frac{(z-\mu_Z)^2}{\epsilon_3} = 1$). Figure 1 shows an example of this connectivity where the Gaussian under analysis presents some overlapped regions with the rest of the Gaussians.

Hence, we propose a convolution between the set of displacements that the Gaussians suffer in the spatial updating d_s, and a Gaussian kernel (GK), thus smoothing the spatial evolution of the foreground Gaussians along the sequence obtaining the set of displacement vectors $\hat{d}_s$.

$$\hat{d}_{s,k} = \sum_{i_1,i_2,i_3}^{N_b} \mathrm{GK}(i_1, i_2, i_3) \cdot d(x + i_1, y + i_2, z + i_3), \qquad (5)$$

where N_b is the neighborhood utilized in the Gaussian k smoothness. Hence, we maintain the consistency of the foreground model, in order to give robustness to the overall system.

Also, in order to achieve a better adaptation of the model into the silhouette of the object, we apply a Gaussian split criterion presented in [6], according to the spatial size of the Gaussian. Gaussians with big area are split into two smaller Gaussians in the direction of the eigenvector associated to the largest eigenvalue ($\lambda_{\max}$).

3. PROJECTING 3D FOREGROUND MODEL TO 2D

The 3D foreground model gathers all the information of the foreground object that we want to segment and reconstruct. In order to use it for 2D foreground segmentation in each view, we need to project the 3D Gaussians to each one of the cameras according to the visibility that the surface voxels present from every view. Hence, a $\mathbb{R}^3 \rightarrow \mathbb{R}^2$ projection is proposed in each camera sensor C_j:

First, a visibility test of the surface voxels is performed for each one the views. We consider only the foreground voxels that are visible from camera C_j thus rejecting all those foreground voxels that appear occluded by the visible ones. The visibility test consists in obtaining the distance from the sensors to each one of the foreground voxels, thus obtaining the minimum distance d_{min} in each projection line corresponding to the closer voxel to the camera. Applying this to each one of the camera sensors, we obtain the bag of visible voxels ν for each view: ν^{C_j}.

Next, we assign each voxel $v_i \in \nu^{C_j}$ to the 3D Gaussian k that maximizes the Equation (4), thus obtaining the group of Gaussians that model visible voxels from each one of the views ζ^{C_j}.

Therefore, for each one of the views C_j, we project the visible Gaussians belonging to ζ^{C_j} according to the projection matrices and focal length that each camera sensor presents. These Gaussians will be used in the 2D planar foreground segmentation for each camera according to [6].

4. RESULTS

We have evaluated our proposal by analyzing four multi-view sequences, of the database presented in [9], which present strong difficulties to achieve a correct 3D reconstruction due to the similarity between some foreground regions and the background. These sequences have been recorded with different acquisition setups in order to better analyze the effect of the errors tolerance in the volumetric reconstruction: These four sequences recorded with 18 cameras (Open arms), 16 cameras (baton and karate) and 8 cameras (dancer). One representative view of the overall multi-view sequence has been selected in each case. In these tests we want to evaluate the viability of the 3D SCGMM to represent the foreground object in the 3-dimensional space, and the subsequent 2-dimensional foreground segmentations that take place in each view by means of the 3D model projection to the 2D images. Hence, we will show in this section qualitative and quantitative results of the current proposal.

For each one of the sequences, the proposed system has been applied in order to obtain the 3D SCGMM of the objects under analysis. The number of Gaussians used in order to form the foreground model in each sequence is closed to 100. Figure 2 displays the 3D spatial representation of the models created in each one of the sequences. We can observe how the combination of each one of the ellipsoids that represents each spatial Gaussian adapts well to the real shape of the objects achieving a complete 3D characterization. Analogously to the 2D SCGMM, the number of Gaussians of the model determines the precision of the modeling: the higher the number of Gaussians of the model, the better the definition of the 3D SCGMM, but the computational cost will increase proportionally. In this evaluation, around one hundred Gaussians have been used for each model in order to achieve a correct characterization of the foreground object.

Qualitative results are displayed in Figure 3 for the *dancer* sequence, where four frames. In second column we can observe the projection of the 3D SCGMM to the view under analysis. Here, the Gaussians of the 3D model are projected to the view only if they model any of the visible voxels obtained for each camera ν^{C_j}. Each Gaussian is drawn with the mean RGB color that each one is modeling, and we can observe how the 2D spatial-color representation adjust correctly to the real shape of the object.

In the third column we can see the 2D foreground segmentation obtained by using the 3D probabilistic model(depicted in sec-

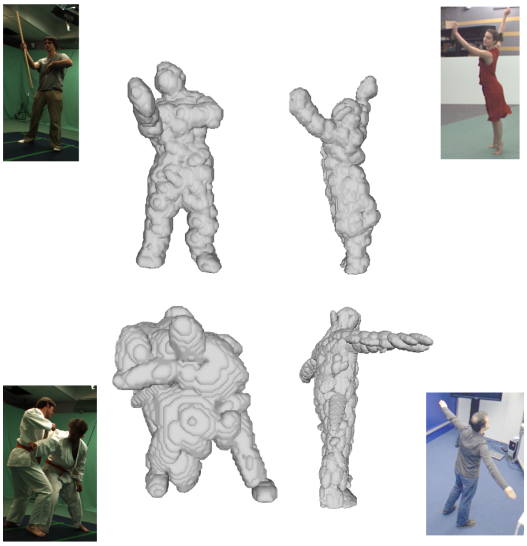


Fig. 2. Resultant foreground 3D SCGMM. Each ellipsoid represents one Gaussian of the foreground model projected to one 2D view.

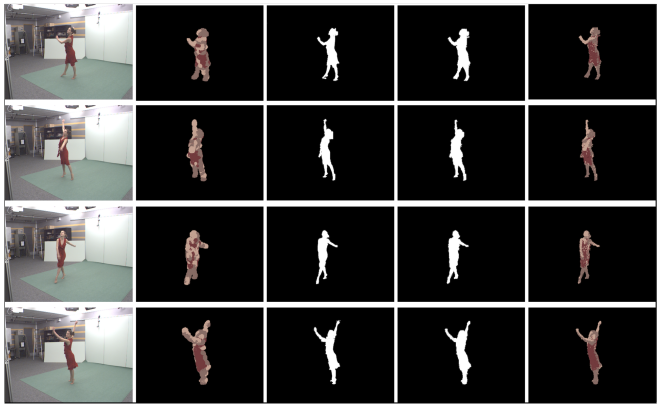
ond column) in the Bayesian MAP-MRF foreground segmentation explained in [6]. This segmentation achieves correct results also in those regions where foreground and background present camouflage situations. The robustness added by the 3D modeling avoids independent 2D errors to be propagated in consecutive frames.

Fourth column shows the 3D volumetric reconstruction with Tolerance to errors $\tau = 1$, computed with all the 2D silhouettes of the multi-view sequence and projected to the views under analysis. We can observe that the final reconstruction presents correct results since we reduce the percentage of errors in the 2D silhouettes. In order to depict the color modeling that the foreground 3D SCGMM is applying to the 3D object, fifth column shows the volumetric reconstruction of the object where each foreground voxel is colored with the RGB color of the Gaussian that better represents it, according to the Equation 4. Hence, we can realize that the 3D SCGMM achieves a correct color-spatial representation of the object along the sequence.

Quantitative results are displayed in Table 1, where the Precision, Recall and f_{measure} results of the frames compared with the ground truth are displayed for each one of these sequences. We can see how the overall results are very similar to the ones obtained by means of method Bayes+sh.rem., since, when the models are correctly initialized, both approaches present similar features. Note that only strong false negative errors in the 3D volumetric reconstruction could lead to errors in the 3D probabilistic modeling, which could propagate the errors to next frames of the sequence, thus producing a degeneration of the 3D SCGMM. More qualitative and quantitative results will be accessible in our web page [1]

Regarding the computational cost, considering a foreground model with no more than 100 Gaussians, we approximate a computational cost of 0.08 frames/second, analyzing a standard sequence with 18 cameras, and using an Intel Core2 Duo 3GHz processor and 20 GB RAM. Since the system proposed performs several time consuming computations at pixel-wise level, like for instance the computation of the likelihoods for each one of the models over every pixel, this computational cost can be improved by developing more efficient algorithms which could work over GPU in a parallel way on these parts of the algorithm.

[1] http://www.jaimegallego.com.es/icip2014_3Dmodel



(a) Dancer sequence. 8 cameras.

Fig. 3. Qualitative results. In the first column, the original frames. Second column shows the 3D SCGMM projection to the view under analysis, where each ellipse represents one Gaussian of the model with the mean color that each one is modeling. Third column is the 2D foreground segmentation obtained by means of the model depicted in second column. Fourth column displays the 3D reconstruction projected to the view under analysis, obtained by means of the foreground segmentation of each view. Fifth column is the 3D reconstruction where each voxel is colored with the mean RGB color value of the 3D Gaussian that better represents the voxel (according to Equation 4).

Table 1. **Quantitative results**

Sequences	Method	Precision	Recall	f_{measure}
Stick	3D SCGMM	$0,98$	0.97	0.98
	Bayes+sh.rem.	$0,97$	0.94	0.96
Dancer	3D SCGMM	0.96	0.96	0.96
	Bayes+sh.rem.	0.94	0.97	0.95
Karate	3D SCGMM	0.97	0.97	0.97
	Bayes+sh.rem.	0.98	0.98	0.98
Open arms	3D SCGMM	0.92	0.97	0.95
	Bayes+sh.rem.	0.95	0.95	0.95

5. CONCLUSIONS

We have presented in this paper a foreground segmentation system for multi-view smart-room scenarios that uses a parametric non-rigid probabilistic model to characterize the object under analysis in the 3D space. This new technique develops a multi-view foreground segmentation system, which combines the information obtained from each one of the views to define the 3D SCGMM for the 3D volumetric representation of the object under analysis. This probabilistic modeling of the object achieves a robust representation of the foreground object, which is projected to each view to perform a Bayesian foreground segmentation [6]. This system achieves correct results, by reducing the false positive and false negative errors in sequences where some camera sensors can present camouflage situations between foreground and background. Finally, we would like to introduce the possibilities that this model could represent in objects recognition or human activity understanding.

6. REFERENCES

[1] L. Bo and C. Sminchisescu. Twin gaussian processes for structured prediction. *Int. Journal of Computer Vision*, 87(1-2):28–52, 2010.

[2] A. Dempster, N. Laird, D. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[3] L. Díaz-Más, R. Muñoz-Salinas, F. Madrid-Cuevas, and R. Medina-Carnicer. Shape from silhouette using dempster–shafer theory. *Pattern Recognition*, 43(6):2119–2131, 2010.

[4] J.-S. Franco and E. Boyer. Fusion of multiview silhouette cues using a space occupancy grid. In *Int. Conf. on Computer Vision*, volume 2, pages 1747–1753. IEEE, 2005.

[5] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture. *International Journal of Computer Vision*, 87(1-2):75–92, 2010.

[6] J. Gallego, M. Pardàs, and G. Haro. Enhanced foreground segmentation and tracking combining bayesian background, shadow and foreground modeling. *Pattern Recognition Letters*, 33(12):1558–1568, 2012.

[7] J. Gallego, J. Salvador, J. R. Casas, and M. Pardas. Joint multi-view foreground segmentation and 3d reconstruction with tolerance loop. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 997–1000. IEEE, 2011.

[8] L. Guan, J.-S. Franco, E. Boyer, and M. Pollefeys. Probabilistic 3d occupancy flow with latent silhouette cues. pages 1379–1386, 2010.

[9] INRIA. 4D Repository. http://4drepository.inrialpes.fr/.

[10] J. Landabaso and M. Pardàs. Cooperative background modelling using multiple cameras towards human detection in smart-rooms. In *In Proc. of European Signal Processing Conference*, 2006.

[11] C.-S. Lee and A. Elgammal. Coupled visual and kinematic manifold models for tracking. *Int. Journal of Computer Vision*, 87(1-2):118–139, 2010.

[12] T. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2-3):90–126, 2006.

[13] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[14] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *Eur. Conf. on Computer Vision*, pages 702–718. Springer, 2000.

[15] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4–27, 2010.

[16] C. Stoll, N. Hasler, J. Gall, H. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *Proc. IEEE Int. Conf. Computer Vision*, pages 951–958, 2011.