Model-based Motion Capture for Crash Test Video Analysis

Juergen Gall[1], Bodo Rosenhahn[1], Stefan Gehrig[2], and Hans-Peter Seidel[1]

[1] Max-Planck-Institute for Computer Science,
Campus E1 4, 66123 Saarbrücken, Germany
[2] Daimler AG, Environment Perception, 71059 Sindelfingen, Germany
{jgall, rosenhahn, hpseidel}@mpi-inf.mpg.de stefan.gehrig@daimler.com

Abstract. In this work, we propose a model-based approach for estimating the 3D position and orientation of a dummy's head for crash test video analysis. Instead of relying on photogrammetric markers which provide only sparse 3D measurements, features present in the texture of the object's surface are used for tracking. In order to handle also small and partially occluded objects, the concepts of region-based and patch-based matching are combined for pose estimation. For a qualitative and quantitative evaluation, the proposed method is applied to two multi-view crash test videos captured by high-speed cameras.

1 Introduction

Fig. 1. Left: *a)* Estimating the pose of the dummy's head from crash test videos is very challenging. The target object is relatively small and partially occluded by the airbag. Furthermore, background clutter and the car's shadow make it difficult to distinguish the head from the background. **Right:** *b)* Estimated pose of the dummy's head. The 3D surface model is projected onto the image.

The analysis of crash test videos is an important task for the automotive industry in order to improve the passive safety components of cars. In particular,

the motion estimation of crash test dummies helps to improve the protection of occupants and pedestrians. The standard techniques for crash analysis use photogrammetric markers that provide only sparse 3D measurements, which do not allow the estimation of the head orientation.

In this work, we address motion capture of rigid body parts in crash test videos where we concentrate on the head – one of the most sensitive body parts in traffic accidents. As shown in Figure 1 a), this is very challenging since the head covers only a small area of the image and large parts are occluded by the airbag. In addition, shadows and background clutter make it difficult to distinguish the target object from the background. To this end, we propose a model-based approach that estimates the absolute 3D rotation and position of the object from multiple views independently of photogrammetric markers. In order to make the estimation robust to occlusions, reference images are synthesized using a 3D model that contains the geometry and the texture of the object. Since our approach further combines region-based and patch-based matching, reliable estimates are obtained even for small body parts as demonstrated in Figure 1.

The 3D surface model of a crash test dummy is readily available as most dummies are manufactured according to ISO standards. The texture is often not provided but it can be acquired by projecting the images from the calibrated cameras on the object's surface using the technique described in [1] to align the 3D model to the images.

2 Related Work

The knowledge of a 3D surface model has been widely used for tracking humans or human body parts, see e.g. [2] or [3]. Besides silhouettes and edges, optical flow is another popular cue for model-based tracking, see e.g. [4]. More recently, segmentation and pose estimation has been coupled [5] where the projected surface of the previous estimated pose serves as a shape prior for the segmentation. It has the advantage that a static background is not required in contrast to methods that rely on background subtraction. Since the performance depends on the accuracy of the shape prior, optical flow can be used to predict the shape of the target object [6]. These approaches are, however, not suitable for crash test videos since they tend to fail in the case of occlusions.

Using patch-based matching, the tracking can be regarded as a detection problem [7] where patches of a textured model from different viewpoints are extracted in a preprocessing step. During tracking, each frame is then matched to one of the keyframes. Although this approach is very fast, it cannot be applied to small and occluded objects where only few features are available. The same problem arises for approaches that combine an iterative analysis-by-synthesis scheme with optical flow [8] or patch-based matching [9]. In order to handle occlusions and small objects, we propose an analysis-by-synthesis approach that combines the concepts of region-based and patch-based matching.

Tracking of small objects for crash video analysis without a surface model has been investigated in [10], where the relative transformation is reconstructed

from a 3D point cloud that is tracked using KLT [11] and stereo depth data. In contrast to model-based approaches, point clouds do not provide all relevant information like depth of penetration or absolute head orientation.

3 System Overview

An outline of our approach is given in Figure 2. Patch-based matching (Section 4.1) is used for pose prediction and for establishing correspondences between the current image and the synthesized reference image. For synthesis, the textured model is projected onto the current image using the predicted pose. Although the reference images provide a relatively small number of matches due to illumination differences, they help to prevent an error accumulation since the static texture of the model is not affected by tracking errors. The small size of the object and temporarily occlusions actually yield very few matches from patches. Additional correspondences are therefore extracted by region matching (Section 4.2) where the segmentation is improved by a shape prior from the predicted pose. The final pose is then estimated from weighted correspondences (Section 4.3) established by prediction, region-based matching, and patch-based matching.

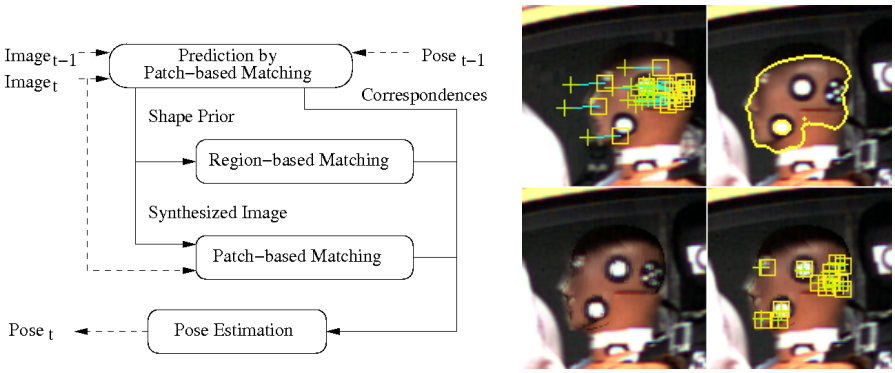

Fig. 2. Left: *a)* Having estimated the pose for time $t-1$, the pose for the next frame is predicted by matching patches between the images of frames $t-1$ and t. The predicted pose provides a shape prior for the region-based matching and defines the pose of the model for synthesis. The final pose for frame t is estimated from weighted correspondences emerging from the prediction, region-based matching, and patch-based matching. **Right:** *b) From top left to bottom right*: Correspondences between two successive frames (*square*: frame $t-1$; *cross*: frame t). Estimated contour. Synthesized image. Correspondences between synthesized image (*square*) and original image (*cross*).

4 Pose Tracking

4.1 Patch-based Matching

Patch-based matching extracts correspondences between two successive frames for prediction and between the current image and a synthesized image for avoiding drift as outlined in Figure 2. For reducing the computation effort of the keypoint extraction [12], a region of interest is selected by determining the bounding box around the projection and adding fixed safety margins that compensate for the movement. As local descriptor for the patches, we apply PCA-SIFT [13] that is trained by building the patch eigenspace from the object texture. 2D-2D correspondences are then established by nearest neighbor distance ratio matching [14]. Since each 2D keypoint x of the projected model is inside or on the border of a triangle with vertices v_1, v_2, and v_3, the 3D counterpart is approximated by $X = \sum_i \alpha_i V_i$ using barycentric coordinates $(\alpha_1, \alpha_2, \alpha_3)$. The corresponding triangle for a 2D point can be efficiently determined by a look-up table containing the color index and vertices for each triangle.

The patch matching produces also outliers that need to be eliminated. In a first coarse filtering step, mismatches are removed by discarding 2D-2D correspondences with an Euclidean distance that is much larger than the average. After deriving the 3D-2D correspondences, the pose is estimated and the new 3D correspondences are projected back. By measuring the distance between the 2D correspondences and their reprojected counterparts, the remaining outliers are detected.

4.2 Region-based Matching

Region-based matching minimizes the difference between the projected surface of the model and the object region extracted in the image, see Figure 2 b). For this purpose, 2D-2D correspondences between the contour of the projected model and the segmented contour are established by a closest point algorithm [15]. Since we are interested in 3D-2D correspondences between the model and the image, we consider only the projected mesh vertices on the model contour where the 3D coordinates are known. The 2D counterpart is then given by the point on the segmented contour that is closest to the projected vertex.

The silhouette of the object is extracted by a level-set segmentation that divides the image into fore- and background where the contour is given by the zero-line of a level-set function Φ. As proposed in [5], the level-set function Φ is obtained by minimizing the energy functional

$$\begin{aligned} E(\Phi) = &- \int_\Omega H(\Phi) \ln p_1 + (1 - H(\Phi)) \ln p_2 \, dx \\ &+ \nu \int_\Omega |\nabla H(\Phi)| \, dx + \lambda \int_\Omega (\Phi - \Phi_0)^2 \, dx, \end{aligned} \tag{1}$$

where H is a regularized version of the step function. The densities of the fore- and background p_1 and p_2 are estimated by a Parzen estimator with Gaussian

kernels. While the first term maximizes the likelihood, the second term regulates the smoothness of the contour by parameter $\nu = 2$. The last term penalizes deviations from the projected surface of the predicted pose Φ_0 where we use the recommended value $\lambda = 0.06$.

4.3 Pose Estimation

For estimating the pose, we seek for the transformation that minimizes the error of given 3D-2D correspondences denoted by pairs (X_i, x_i) of homogeneous coordinates. To this end, we represent a 3D rigid motion M by a twist $\theta\hat{\xi}$ [4]: $M = \exp(\theta\hat{\xi})$. Hence, a transformation of a point X_i is given by

$$X_i' = \exp(\theta\hat{\xi})X_i. \tag{2}$$

Since each 2D point x_i defines a projection ray that can be represented as Plücker line $L_i = (n_i, m_i)$ [16], the error of a pair (X_i', x_i) is given by the norm of the perpendicular vector between the line L_i and the point X_i'

$$\left\| \Pi\left(X_i'\right) \times n_i - m_i \right\|_2, \tag{3}$$

where Π denotes the projection from homogeneous coordinates to non-homogeneous coordinates. Using the Taylor approximation $\exp(\theta\hat{\xi}) \approx I + \theta\hat{\xi}$ where I denotes the identity matrix, Equation (2) can be linearized. Hence, the sought transformation is obtained by solving the weighted linear least squares problem

$$\frac{1}{2}\sum_i w_i \left\| \Pi\left(\left(I + \theta\hat{\xi}\right) X_i\right) \times n_i - m_i \right\|_2^2, \tag{4}$$

i.e. by solving a system of linear equations.

For estimating the final pose, correspondences from the prediction (C_p), region-based matching (C_r), and the synthetic image (C_s) are used as outlined in Figure 2. Since the number of correspondences from region matching varies according to scale, shape, and triangulation of the object, we weight the summands in Equation (4) such that the influence between patches and silhouette is independent of the model. This is achieved by setting the weights for the equations for C_p and C_s in relation to C_r:

$$w_p = \alpha\frac{|C_r|}{|C_p|}, \quad w_r = 1, \quad w_s = \beta\, w_p. \tag{5}$$

While the influence of the image-based patches and the contour is controlled by the parameter α independently of the number of correspondences, the weight w_s reflects the confidence in the matched patches between the synthesized and original image that increases with the number of matches $|C_s|$ relative to $|C_p|$. Since illumination differences between the two images entail that $|C_s|$ is usually less than $|C_p|$, the scaling factor β compensates for the difference. In our experiments, we have obtained good results with $\alpha = 2.0$ and $\beta = 2.0$.

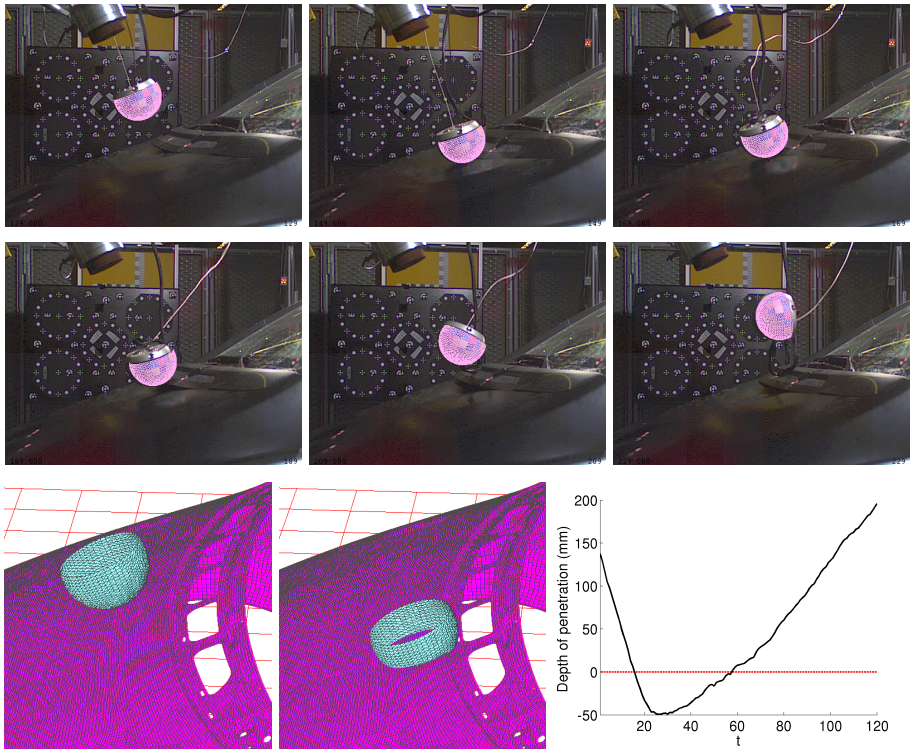

Fig. 3. Rows 1, 2: The head crashes onto the engine hood. Estimates for frames 5, 25, 45, 65, 85, and 105 are shown (*from top left to bottom right*). The pose of the head is well estimated for the entire sequence. **Row 3:** Virtual reconstruction of the crash showing the 3D surface model of the head and of the engine hood. **From left to right:** *a)* Frame 5. *b)* Frame 25. The head penetrates the engine hood. *c)* Depth of penetration. The black curve shows the distance of the head to the engine hood (*dashed line*).

5 Experiments

The first experiment investigates the dynamics of a pedestrian head crashing onto the engine hood, see Figure 3. The sequence has been captured at 1000 Hz by two calibrated cameras with 512×384 pixel resolution. For segmentation, the images have been converted to the CIELab color space that mimics the human perception of color differences. Since we have registered the engine hood as shown in row 3 of Figure 3, the depth of penetration can be measured from the estimated head pose. In this case, the head penetrates $49.1mm$ into the engine compartment. This is a relevant information for crash test analysis since severe head injuries might be caused by crashing into the solid engine block. Note that a standard silhouette-based approach would not be able to estimate the rotation

due to the symmetric shape of the object whereas our approach provides good results for the entire sequence.

For the second experiment, the head of a dummy is tracked during a EuroNCAP offset crash where the car drives into an aluminum barrier with 40% overlap as shown in Figure 4. Due to the barrier, the car jumps and moves laterally. Although the sequence was captured at 1000 Hz by 3 cameras with 1504×1128 pixel resolution, the head covers only 70×70 pixels, i.e., less than 0.3% of the image pixels. In addition, the head is occluded by more than 50% at the moment of the deepest airbag penetration, and the segmentation is hindered by shadows and background clutter. Nevertheless, Figure 7 demonstrates that the head pose is well estimated by our model-based approach during the crash. The trajectory in Figure 5 reflects the upward and lateral movement of the car away from the camera due to the offset barrier. For a quantitative error analysis, we

Fig. 4. Three frames of the EuroNCAP offset crash sequence. The car jumps and moves laterally due to the offset barrier. The head is occluded by more than 50% at the moment of the deepest airbag penetration.

have compared the results with a marker-based system using photogrammetric markers. The 3D tracking error is obtained by the Euclidean distance between the estimated position and the true position of the 5-dot marker on the left hand side of the dummy's head. The results are plotted in Figure 5 where the average error is $37mm$ with standard deviation of $15mm$. For computing the velocity and the acceleration of the head, the trajectories from the marker-based and the model-based method are slightly smoothed, as it is common for crash test analysis. Figure 6 shows that the velocity and the acceleration are well approximated by our approach. A comparison with an acceleration sensor attached to the head further reveals that the deceleration is similar to the estimates of our approach. For the offset crash sequence, our current implementation requires 6 seconds per frame on a consumer PC.

Finally, we remark that the marker-based system provides only one 3D point for the position of the head whereas our approach estimates the full head pose. It allows additional measurements like rotation of the head or penetration depth which help to analyze crash test videos.

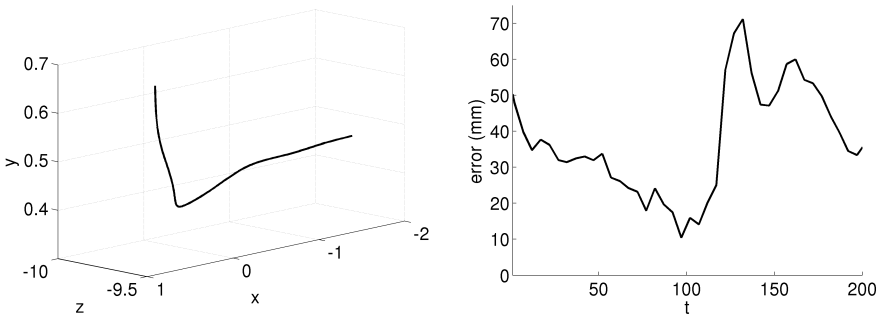

Fig. 5. Left: 3D trajectory of the head. **Right:** 3D tracking error of the head. The ground truth is obtained from a marker-based system with standard deviation $\pm 2.5mm$ for the x- and y-coordinates and $\pm 5mm$ for the z-coordinate. Note that the object is about $10m$ away from the camera.

6 Conclusion

We have presented a model-based approach for pose tracking of rigid objects that is able to meet the challenges of crash test analysis. It combines the complementary concepts of region-based and patch-based matching in order to deal with the small size of the objects. Since the targets are temporarily occluded, we have proposed the use of synthesized reference images, which help to avoid drift and to recover from prediction errors. In contrast to conventional marker-based systems, our approach estimates all six degrees of freedom of dummy body parts like the head. This opens up new opportunities for analyzing pedestrian crashes where many biomechanical effects are not fully understood. The accuracy and robustness of our system has been demonstrated by a offset crash sequence where a quantitative comparison with a marker-based system and an acceleration sensor is provided.

Acknowledgments. The research was partially funded by the Max Planck Center Visual Computing and Communication and the Cluster of Excellence on Multimodal Computing and Interaction.

References

1. Gall, J., Rosenhahn, B., Seidel, H.P.: Clustered stochastic optimization for object recognition and pose estimation. In: Patt. Recog. Volume 4713 of LNCS., Springer (2007) 32–41
2. Hogg, D.: Model-based vision: A program to see a walking person. Image and Vision Computing **1**(1) (1983) 5–20
3. Gavrila, D., Davis, L.: 3-d model-based tracking of humans in action: a multi-view approach. In: IEEE Conf. on Comp. Vision and Patt. Recog. (1996) 73–80

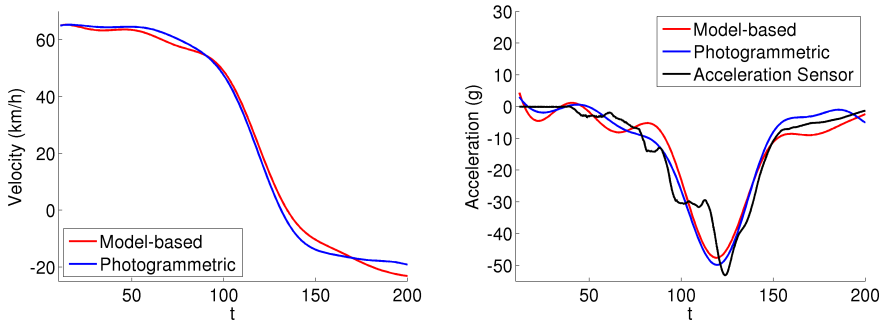

Fig. 6. Comparison with a marker-based system and an acceleration sensor. The model-based approach provides accurate estimates for velocity and acceleration. **Left:** Velocity (x-axis). **Right:** Acceleration (x-axis).

4. Bregler, C., Malik, J., Pullen, K.: Twist based acquisition and tracking of animal and human kinematics. Int. J. of Computer Vision **56**(3) (2004) 179–194
5. Rosenhahn, B., Brox, T., Weickert, J.: Three-dimensional shape knowledge for joint image segmentation and pose tracking. Int. Journal of Computer Vision **73**(3) (2007) 243–262
6. Brox, T., Rosenhahn, B., Cremers, D., Seidel, H.P.: High accuracy optical flow serves 3-d pose tracking: Exploiting contour and flow based constraints. In: European Conf. on Comp. Vision. Volume 3952 of LNCS., Springer (2006) 98–111
7. Lepetit, V., Pilet, J., Fua, P.: Point matching as a classification problem for fast and robust object pose estimation. In: IEEE Conf. on Computer Vision and Patt. Recognition. Volume 2. (2004) 244–250
8. Li, H., Roivainen, P., Forcheimer, R.: 3-d motion estimation in model-based facial image coding. IEEE Trans. Pattern Anal. Mach. Intell. **15**(6) (1993)
9. Gall, J., Rosenhahn, B., Seidel, H.P.: Robust pose estimation with 3d textured models. In: IEEE Pacific-Rim Symposium on Image and Video Technology. Volume 4319 of LNCS., Springer (2006) 84–95
10. Gehrig, S., Badino, H., Paysan, P.: Accurate and model-free pose estimation of small objects for crash video analysis. In: Britsh Machine Vision Conference. (2006)
11. Shi, J., Tomasi, C.: Good features to track. In: IEEE Conf. on Comp. Vision and Patt. Recog. (1994) 593–600
12. Lowe, D.: Object recognition from local scale-invariant features. In: Int. Conf. on Computer Vision. (1999) 1150–1157
13. Ke, Y., Sukthankar, R.: Pca-sift: A more distinctive representation for local image descriptors. In: IEEE Conf. on Comp. Vision and Patt. Recog. Volume 2. (2004) 506–513
14. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Conf. on Computer Vision and Patt. Recognition **02** (2003) 257–263
15. Zhang, Z.: Iterative point matching for registration of free-form curves and surfaces. Int. Journal of Computer Vision **13**(2) (1994) 119–152
16. Stolfi, J.: Oriented Projective Geometry: A Framework for Geometric Computation. Academic Press, Boston (1991)

Fig. 7. Estimated pose of the dummy's head for frames 7, 22, 37, 52, 67, 82, 97, 112, 127, 142, 157, 172, 187, 202, and 217 (*from top left to bottom right*).