Multi-view Based Estimation of Human Upper-Body Orientation

Lukas Rybok*, Michael Voit†, Hazım Kemal Ekenel* and Rainer Stiefelhagen*†
* *Institute of Anthropomatics,*
Karlsruhe Institute of Technology, Karlsruhe, Germany
Email: {rybok, ekenel, rainer.stiefelhagen}@kit.edu
† *Interactive Analysis and Diagnosis,*
Fraunhofer IOSB, Karlsruhe, Germany
Email: michael.voit@iosb.fraunhofer.de

***Abstract*—The knowledge about the body orientation of humans can improve speed and performance of many service components of a smart-room. Since many of such components run in parallel, an estimator to acquire this knowledge needs a very low computational complexity. In this paper we address these two points with a fast and efficient algorithm using the smart-room's multiple camera output. The estimation is based on silhouette information only and is performed for each camera view separately. The single view results are fused within a Bayesian filter framework. We evaluate our system on a subset of videos from the CLEAR 2007 dataset [1] and achieve an average correct classification rate of 87.8 %, while the estimation itself just takes 12 ms when four cameras are used.**

***Keywords*-body-orientation; multi-view fusion; bag-of-features**

I. Introduction and Related Work

In recent years much research has been conducted towards enhancing the way people interact by unobtrusively providing services in a smart-room environment. CHIL (http://chil.server.de) and AMI (http://amiproject.org) are just two of the major European projects in this field. A typical smart-room is equipped with multiple acoustic and visual sensors which are used to analyze the scene in order to both provide the user contextual information and attempt to anticipate the user's needs. Therefore, an essential role in a smart-room architecture is played by components providing means to percept human activities.

The knowledge about the orientation of the body (angle around the axis perpendicular to the ground plane) can be a useful cue. While tracking the position of people, this information can help to deduce the most likely movement direction of the target. Model-based human motion capture systems can also benefit from this as the process of model-fitting is computationally very costly and the incorporation of the body orientation can reduce its complexity.

Since in many cases the legs can be occluded by furniture or other objects, it is beneficial to estimate the orientation of the upper-body only. The estimator should also yield a very low computational complexity to be useful because it would be just one of many software components running in parallel. Thus, we present in this paper a fast approach to classify the upper-body orientation of a person inside of a smart-room.

To our best knowledge not much research has been conducted to explicitly estimate the orientation of the human body. In [2] the estimation is based on the motion of a tracked person and the size of its bounding ellipse. An analysis of the shape of the silhouette in images from ceiling mounted cameras is performed in [3], [4]. These approaches, though very fast, just give a course estimation and are sensitive to arm movement. Peng and Quian [5] use a multi-camera scenario and perform multi-linear analysis to extract orientation vectors from binary silhouette images. From these a one-dimensional manifold is learned which is used to retrieve the body orientation by solving a nonlinear least squares problem. However, a disadvantage of this approach is its low speed.

In contrast to other works, our system is fast, accurate, and also flexible in regard to the number of used camera views at the same time. To be independent of the user's appearance, we estimate the upper-body orientation using silhouette information only. This approach however introduces ambiguities that are difficult to resolve using a monocular setup (see Fig. 1). Therefore we show in our work that by increasing the number of camera views such ambiguities can be partly compensated and an overall increase in the system performance is achieved. In particular, we investigate the encoding of the silhouette information using either shape contexts or histogram of shape context (*HoSC*) descriptors as a base for the estimation. For each camera view in the smart-room the orientation angle is classified separately and the single hypotheses are fused within a Bayesian filter framework. This way we only need to train one estimator that can be used for any view as long as the inclination angle of the camera to the ground plane is similar with the views used for training. We evaluated our approach on five sequences from the CLEAR 2007 evaluation dataset and achieved an average correct classification rate of 87.8%, when using 12 orientation classes.

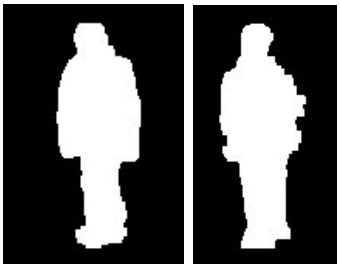

Figure 1. Ambiguity when only using silhouette based features: it is difficult to tell appart a frontal view (left) from a rear view (right).

II. Estimating the Upper-Body Orientation

Our approach is based on silhouette features as they are easy to extract by using foreground-background segmentation, but at the same time contain much relevant information to estimate the upper body orientation angle. For each camera view the estimation is carried out separately. The upper-body region is extracted and the silhouette is encoded using local features described in Sec. II-A. We then classify the extracted feature vector to obtain a hypothesis for the orientation angle based on the observation of each camera and fuse the results within a Bayesian filter framework as described in Sec. II-B. Since we want to omit the need of training a separate classifier for each view, the output of the classifiers is given relative to the coordinate system of the capturing camera. The transformation to the world coordinate system is performed while merging the single hypotheses.

A. Image Descriptors

Since segmentation algorithms are not reliable enough to provide noise free silhouettes, we need an encoding of the silhouette information that is partly robust to segmentation failures. The image descriptor should also be invariant to scaling as the input silhouettes are usually not of the same size. Both shape contexts and HoSC descriptors provide these advantages.

In [6] shape contexts are introduced as rich local descriptors of contour information. In order to transform a shape to its shape context representation, the silhouette points are sampled, resulting in a set of n points. For each sampled point a shape context can be calculated as a histogram of the relative position of its neighboring points. The histogram bins are uniform in log-polar space making shape contexts more sensitive to differences of nearby points. Invariance to scaling of the shapes is obtained by normalizing the point distances with the mean distance between all point pairs. As the histograms are calculated relative to each point, shape contexts are also translation invariant.

A major disadvantage of shape context descriptors is its high dimensionality. For instance, the parameters used in our experiments lead to 2400-dimensional feature vectors. A dimensionality reduction can be achieved using a bag-of-features scheme as described in [7]. The space spanned by all shape contexts obtained from the training images is first vector quantized using k-means clustering which yields a k-dimensional codebook of the cluster means. In order to transform a new sample shape to a HoSC feature, its shape contexts are binned to a histogram where each bin is associated to one of the codebook vectors. Effects of spatial quantization are reduced by employing a soft-voting scheme. Thus we calculate the contribution η of a shape context $\vec{SC}$ to a cluster $\vec{C}_i$ as in [8] using:

$$\eta_i(\vec{SC}) = \frac{min_{j=1\ldots k}|\vec{SC} - \vec{C}_j|^2}{|\vec{SC} - \vec{C}_i|^2} \tag{1}$$

The obtained η_i are summed up over all shape contexts and the resulting histogram is normalized to unit length.

B. Multi-view Fusion

Similar to our previous work [9], we merge the estimations that are performed for each view to obtain the final orientation class $\hat{\theta}_t$ by using Bayes' theorem:

$$\hat{\theta}_t = \arg\max_{x_i \in X} c \cdot p(Z_t|x_i)P(x_i) \tag{2}$$

Here $p(Z_t|x_i)$ denotes the class-conditional probability of an observation Z_t at time t given the i-th orientation class x_i, with $X = \{x_i\}$ being the discrete state-space. $P(x_i)$ is the prior probability of the occurrence of class x_i and c is a normalizing constant being equal to the total probability. The class-conditional $p(Z_t|x_i)$ is calculated by averaging the classification confidence for each view, which is based on a confusion matrix learned on validation data. The prior $P(x_i)$ depends on the previous state x' at time $t-1$, as the turning speed of the body is usually limited. Thus it can be computed as:

$$P(x_i) = \sum_{x' \in X} P(x_i|x')P(x'|Z_{t-1}), \tag{3}$$

where the state-transition probability $P(x_i|x')$ we model with a zero-mean Gaussian $N_{0;\sigma}(x_i - x')$. We determined the standard deviation σ during experiments on validation data and set it to $20°$. When no preceding state is known, we assume the prior to be equal for each state.

III. Experiments

We evaluated the proposed approach on a subset of video sequences from the CLEAR 2007 head-pose evaluation dataset. Each of the sequences consists of about 2700 frames and contains one person that randomly changes his body orientation. The camera setup used to capture the videos can be seen in Fig. 2. Within the sequences we annotated the position of the shoulders in all views, so that the body

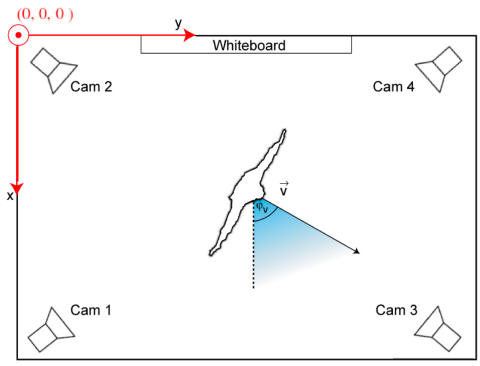

Figure 2. Camera setup of the smart-room

Table I
OVERALL PERFORMANCE OF THE ALL-TO-ALL CLASSIFICATION SCHEME

Method	Correct Class
hosc_svm	82.6 %
hosc_nmc	79.1 %
sc_nmc	85.0 %

Table II
OVERALL PERFORMANCE OF THE ONE-TO-ONE CLASSIFICATION SCHEME

Method	Correct Class
hosc_svm	65.9 %
hosc_nmc	76.8 %
sc_nmc	87.8 %

position and orientation could be retrieved using multi-view geometry. Additionally, we segmented foreground regions using the algorithm presented in [10]. Because the annotation process is very time-consuming, our test set consisted of five 3 minutes long sequences (15 fps, 320×240 pixels), each containing one different person.

The evaluation was performed using a leave-one-out approach, where each sequence was used as testing data once, while the remaining sequences were used for training. In our experiments we evaluated SVM classification together with HoSC features (*hosc_svm*) and Nearest-Mean classification (NMC) with both shape contexts (*sc_nmc*) and HoSC features (*hosc_nmc*). Because of the high dimensionality of shape context features and limited amount of training data, we did not further investigate SVM classification of shape contexts in our experiments. This was also supported by very low initial experimental results. We favor the use of NMC over Nearest-Neighbor classification, since the low speed of the latter one prohibits its use in a smart-room environment.

For the SVM classification we use a RBF kernel and a one-vs-one approach. We handle the multi-class case with a max-win voting strategy, where each SVM votes for one class and the recognition result is the mostly voted class. Both classifiers use the χ^2 metric as a distance function between features. When encoding a silhouette with shape contexts, all extracted features from one silhouette are concatenated to one feature vector. We determined all parameters experimentally on separate validation data and set the histogram dimensions to $r = 6$ and $a = 8$ for HoSC features and $r = 4$ and $a = 8$ for shape contexts. We sampled 50 points equidistantly on the upper-body silhouette and used a 50 element HoSC codebook. The size of one class was set to 30° since it makes the evaluation independent of errors introduced through the labeling process.

In our first experiment we used an *all-to-all* classification scheme where only one classifier is trained using data from all four views and used to estimate the orientation angle in each view. As can be seen in Tab. I, *sc_nmc* outperforms the other approaches with an average correct classification rate of 85.0%.

In a multi-camera setup, the body silhouette my look different depending on the distance from the camera. In order to investigate the impact of such appearance changes on the performance of our approach, we used in the next experiment for each camera a separate classifier trained only with data captured with that camera (*one-to-one* scheme). In Tab. II it can be observed that the performance of the NMC approach deviates only insignificantly from the results in the previous experiment showing its view-independence. However, when using the SVM classifier a clear decrease in performance is visible which we believe is due to the smaller amount of training data per estimator caused by the experimental setup.

In a final experiment, we show the benefits of using a multi-camera setup to estimate the orientation angle. To do this, we evaluated the system performance when using all possible combinations of camera views during testing and averaging the results for each number of views that are used to derive the final estimation. In Fig. 3, it can be clearly seen that with each additional view the performance of all systems improves greatly. When comparing the results of a single camera based estimation with the results of using all four views, the highest relative performance gain of 38% can be observed for *sc_nmc*. These results strongly suggest that basing the estimation on multiple views can compensate for the loss of discrimination between frontal and rear views.

Using an unoptimized C++ implementation on a 3GHz Intel Pentium IV machine, the computation of one concatenated shape context feature vector takes 2ms and its classification with the NMC additional 1ms. This clearly shows the suitability of the proposed approach as an auxiliary component in a smart-room environment.

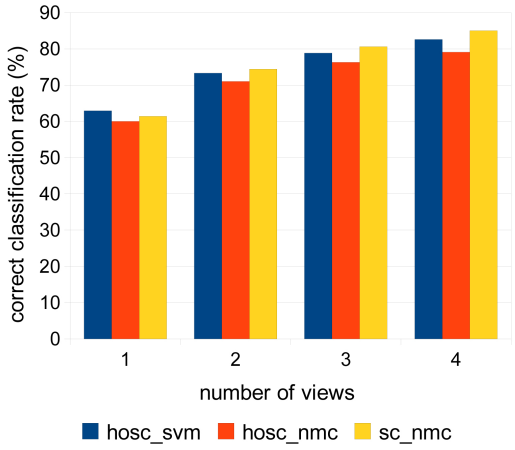

Figure 3. Overall performance (one-to-all scheme) when changing the number of views

IV. Conclusion

We presented a classification based approach to estimate the upper-body orientation of a person inside of a smart-room. For each of the camera views used in the sensor setup of the room, we build a single hypothesis of the orientation angle using silhouette information only. A joint measurement is then obtained with a Bayesian filter framework. This makes our approach flexible, as no retraining is needed whenever the camera setup is changed. We showed that the use of four cameras instead of a single one greatly improves the performance of our approach and achieved an average correct classification rate of up to 87.8% on five sequences from the CLEAR 2007 evaluation dataset. The high performance and low computational complexity (12ms per estimation when using four views) makes our approach well suitable as an auxiliary component in a smart-room scenario.

Acknowledgment

This study is partially funded by the German Research Foundation (DFG) under Sonderforschungsbereich SFB 588 - Humanoid Robots - and by BMBF, German Federal Ministry of Education and Research as part of the GEMS programme.

References

[1] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo, "The CLEAR 2007 evaluation," *CLEAR 2007 and RT 2007*, pp. 3–34, 2008.

[2] M. W. Lee and R. Nevatia, "Body part detection for human pose estimation and tracking," in *IEEE Workshop on Motion and Video Computing*, 2007, p. 23.

[3] S. Iwasawa, J. Ohya, K. Takahashi, T. Sakaguchi, K. Ebihara, and S. Morishima, "Human body postures from trinocular camera images," in *Intl. Conf. on Automatic Face and Gesture Recognition*, 2000, pp. 326–331.

[4] W. Zhang, T. Matsumoto, J. Liu, M. Chu, and B. Begole, "An intelligent fitting room using multi-camera perception," in *Intl. Conf. on Intelligent User Interfaces*, 2008, pp. 60–69.

[5] B. Peng and G. Qian, "Binocular dance pose recognition and body orientation estimation via multilinear analysis," in *CVPR Workshops 2008*, 2008, pp. 1–8.

[6] G. Mori, S. Belongie, and J. Malik, "Shape contexts enable efficient retrieval of similar shapes," in *CVPR 2001*, vol. 1, 2001, pp. 723–730.

[7] A. Agarwal and B. Triggs, "3D human pose from silhouettes by relevance vector regression," in *CVPR*, 2004.

[8] R. Poppe and M. Poel, "Comparison of silhouette shape descriptors for example-based human pose recovery," in *Intl. Conf. on Automatic Face and Gesture Recognition*, 2006, pp. 541–546.

[9] M. Voit, K. Nickel, and R. Stiefelhagen, "A bayesian approach for multi-view head pose estimation," in *IEEE MFI 2006*, 2006, pp. 31–34.

[10] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, 2006.