Localized fusion of Shape and Appearance features for 3D Human Pose Estimation

S.Sedai
suman@csse.uwa.edu.au
M.Bennamoun
m.bennamou@csse.uwa.edu.au
D.Q. Huynh
du@csse.uwa.edu.au

School of Computer Science and Software Engineering
The University of Western Australia
Crawley, Perth
Western Australia 6009

Image based 3D human pose estimation is an important research area because of its application in many different fields such as content-based image retrieval, human computer interaction (HCI), human gait analysis, computer animation in movies, and visual surveillance. There are two basic types of approaches to 3D pose estimation from images. The first type is the *generative method*, where the pose that best explains the observed image features is determined from a population of hypothesized poses, e.g. [2, 3]. The second type is the *discriminative method*, where a direct mapping between image features and pose parameters is learned using training data. Approaches of the first type are, in general, inefficient as the correct 3D pose must be searched in a high dimensional pose space. Discriminative pose estimation approaches such as [1, 6], on the other hand, use only single type of features. In this paper, we present a discriminative approach that fuses shape- and appearance-based features in a learning framework for 3D human pose estimation. To the best of our knowledge, the combination of these two type of features by a discriminative method has not been explored in the context of 3D human pose estimation.

Our method is based on clustering the 3D pose space into several modular regions and learning the regressors for both feature types and their optimal fusion scenario in each region. The optimal fusion scenario is learned using a confidence measure of each feature type in a region. The concept is intuitive because parts of the data space may prefer shape features over appearance features whereas some other parts may prefer the opposite. This way the complementary information of the individual feature types is exploited, leading to an improved performance of the pose estimation system.

We choose to combine the shape- and appearance-based feature types because they have the tendency to complement each other. Shape-based features such as silhouettes are insensitive to background variations. However, one silhouette can be associated with more than one pose, resulting in ambiguities. Such ambiguities are resolved using temporal constancy [6]. Appearance-based features, e.g image textures, are more discriminative than shape features; however, they are unstable due to background clutters and variations in the clothing of the human subject. Therefore, a proper combination of such feature types can improve the performance of the pose estimation system. We use the *Histogram of Shape Context* (HoSC) descriptor [1] constructed from a human silhouette for shape feature presentation. To represent appearance, we use the *Histogram of Local Appearance Context* (HLAC) descriptor [4] based on the gradient orientation histogram computed on the image window containing the human subject. The prototypes of both descriptors along with the block diagram of our fusion system are shown in Figure 1. The relation between the two feature types $\mathbf{x}_1$ and $\mathbf{x}_2$ and the 3D pose $\mathbf{y}$ is approximated by the following fusion model,

$$\mathbf{y} = \sum_{i=1}^{K} g_k(\chi) f_k(\mathbf{x}_1, \mathbf{x}_2), \tag{1}$$

where K is the number of clusters the pose space is partitioned into and χ is an m-dimensional vector obtained by concatenating $\mathbf{x}_1$ and $\mathbf{x}_2$. The partition function $g_k(\chi)$ is a multi-class classifier which gives the probability that the k^{th} cluster is selected for a given feature instance. In each clustered region, the fusion of shape and appearance regressors is then performed locally as follows

$$f_k(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{2} \beta_{kj}(\chi) f_{kj}(\mathbf{x}_j), \tag{2}$$

where f_{kj} is a linear regressor that takes a feature vector $\mathbf{x}_j$ and outputs the 3D pose estimate for the k^{th} cluster; β_{kj} is a combining function which outputs a weight that is proportional to the confidence measure of the regressor f_{kj}. Since we are combining only two feature types, the combiner function is approximated by a two-class classifier using logistic regressor. Hence our fusion model consists of $2K$ regressors, one K -class classifier and K two-class classifiers. We model the regressors and the classifiers using Relevance Vector Machine (RVM) [7] as it is known to give a sparse model while maintaining good generalization performance. The details of the training process to estimate the parameters of the fusion model are described in this paper.

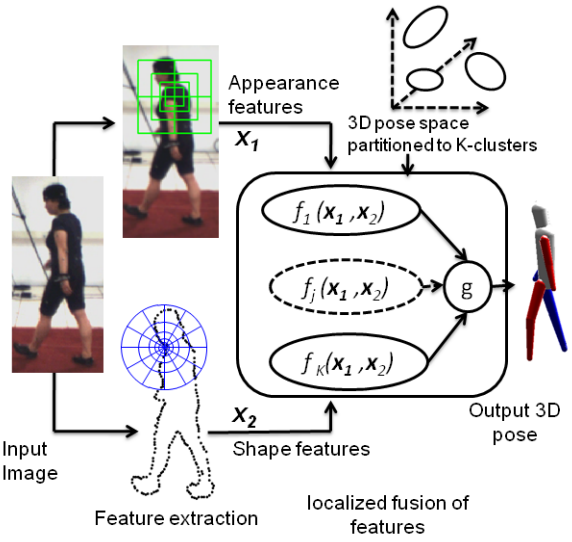

Figure 1: A block diagram of the pose estimation system using localized fusion of shape [1] and appearance [4] feature types. The pose space is divided into several clusters using *kmeans*. For each cluster, a local fusion model of shape and appearance feature type is trained.

We train and evaluate the fusion model using the HumanEva I dataset [5]. Our fusion method produces an average RMS error of 26.96 mm over the test set of the Walking sequence from camera C1. This is an improvement compared to the average RMS errors of 39.58 using shape features and 32.22 mm using appearance features. We found similar improvements on other action sequences (Jogging and Boxing) and other camera views (C2 and C3).

[1] Ankur Agarwal and Bill Triggs. Recovering 3D human pose from monocular images. *IEEE Transactions on PAMI*, 28(1), 2006.

[2] J. Deutscher and I. D. Reid. Articulated body motion capture by stochastic search. *Int'l Journal of Computer Vision*, 61(2), 2005.

[3] Mun Wai Lee and I. Cohen. A model-based approach for estimating human 3D poses in static images. *IEEE Transactions on PAMI*, 28 (6):905–916, 2006.

[4] Suman Sedai, Mohammed Bennamoun, and Du Huynh. Context-based appearance descriptor for 3D human pose estimation from monocular images. In *DICTA*, pages 484–491, 2009.

[5] Leonid Sigal and Michael J. Black. HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, 2006.

[6] Cristian Sminchisescu, Atul Kanaujia, Zhiguo Li, and Dimitris Metaxas. Discriminative density propagation for 3D human motion estimation. In *IEEE Conference on CVPR*, pages 390–397, 2005.

[7] Michael E. Tipping. The relevance vector machine. *Advances in NIPS*, 12:652–658, 2000.