社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－CVIM－155 (12)
2006／9／8

検索－組合せ法を用いた単眼画像からの三次元人物姿勢推定

曹 暉　大西 昇　竹内 義則†　松本 哲也　工藤 博章

名古屋大学大学院 情報科学研究科　†名古屋大学 情報セキュリティ対策推進室
〒464–8603 愛知県名古屋市千種区不老町
E-mail: souki@ohnishi.m.is.nagoya-u.ac.jp

あらまし 本研究では，単眼画像からの三次元人物姿勢の推定を目的とする．事例法とパーツ法を統合して新しい検索－組合せ手法を提案した．全身姿勢に基づく事例データベースの代わりに，簡潔な上下半身姿勢からなる事例データベースを使用する．上下半身姿勢の有効な組合せによって何百万もの全身姿勢を復元することができる．最初に，上下半身姿勢の候補は，事例データベースから，部分的画像照合によって検索される．これらの半身姿勢候補から，事前学習された組合せの制約条件に従って，現実的な組合せが選択される．それから，coarse-to-fine 評価方法によって最適な半身姿勢の組合せを選ぶ．提案手法は，(全身に基づく) 事例法より低い時間／空間複雑さを持って，パーツ法より現実的な三次元姿勢を保証できる．

Retrieval-Combination Approach to Estimating 3D Human Pose from A Monocular Image

Hui Cao　Noboru Ohnishi　Yoshinori Takeuchi†　Tetsuya Matsumoto　Hiroaki Kudo

Graduate School of Information Science, †Infomation Security Promotion Agency, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464–8603 Japan

Abstract The objective of this work is to estimate 3D human pose from a monocular image. We propose a novel retrieval-combination approach that exploits the wide capability of example-based approaches and the flexibility of parts-based approaches. Instead of storing and searching for similar full-body examples, we adopt a half-body representation (i.e., upper-body and lower-body) to reduce a large full-body database into a compact half-body database that has good generalization ability to recover millions of poses by valid half-body combinations. For a given input image, half-body candidates are first retrieved from databases by partial shape matching. Valid half-body combinations of these candidates are selected based on a learned combination constraint, and then we choose the optimal combination(s) in a coarse-to-fine evaluation method. We show good experimental results of pose estimation for both synthetic and real images. Our approach has lower computational and space complexities than example-based approaches and ensures better realistic 3D pose estimates than parts-based approaches.

1 Introduction

Estimating a human body pose from a monocular image is important for such image understanding applications as recognition of human activities, markerless motion capturing, and virtual reality applications. However, recovering a 3D human pose is a challenging problem due to the high dimensionality of state space, parts occlusion, clothes variation, unknown body orientation and so on. Existing approaches to this problem can be categorized into three types: *Parts-based approaches* search for possible limb poses and find the optimal com-

bination(s) of parts based on constraints of kinematic relations between body parts [3]. *Learning-based approaches* learn a model that directly infers poses from observable image quantities [5]. *Example-based approaches* store a set of training examples whose 3D poses are known and estimate poses by searching for training images that resemble input images [6].

However, most existing approaches are not appropriate for the estimation of a wide range of 3D human poses. Learning-based approaches can only deal with a limited set of typical human poses. Parts-based approaches are able to deal with large pose variations; however, the (local) kinematic constraints between adjacent body parts cannot ensure that limb poses will be combined into a (globally) realistic 3D pose. Example-based approaches may be a good choice when millions of examples are available, but high computational and space complexities restrict their use.

In this work we propose a retrieval-combination approach that exploits the extensive capability of example-based approaches and the flexibility of parts-based approaches. Instead of storing and searching for similar *full-body* examples, we use a *half-body* representation (i.e., upper-body and lower-body) to reduce a large full-body database into a compact half-body database that has good generalization ability for recovering millions of poses by valid half-body combinations. Since any half-body pose is actually a valid combination of parts, the problem of invalid combinations of parts that occurred in parts-based approaches can be largely avoided. On the other hand, by using a half-body representation and an efficient retrieval-combination strategy, our approach outperforms examples-based approaches in terms of lower computational and space complexities.

The rest of this paper is organized as follows. Section 2 gives an overview of our approach. From Sections 3 to 6, we detail the main steps of our approach: half-body retrieval, invalid combination pruning, and coarse and fine evaluations. Section 7 describes the experiments performed and discusses our results. Finally, Section 8 contains a conclusion and future research directions.

2 Overview of Our Approach

We represent 3D human poses by 51D vectors including 3D locations for each of 17 key body joints. The input image is a human silhouette. For a given input image, we estimate the 3D human pose by efficiently looking for optimal half-body combination(s) from a database of half-body poses.

The query image is first processed to generate a set of data including a normalized silhouette, contour, Distance Transform (DT), and image features (i.e., Hu Moment). With these data as input, we estimate 3D human poses in the following four steps:

Half-body Pose Retrieval Compute chamfer distance from every upper-body contour in the database to the input contour, and sort all database contours in ascending order of computed distances. The top k ranked contours are marked as upper-body candidates. Similarly k lower-body candidates also are retrieved.

Invalid Half-body Combinations Pruning Among all possible combinations from retrieved half-body candidates, prune invalid ones by checking whether they satisfy the learned constraints of half-body combinations.

Coarse Evaluation of Valid Combinations Sort the remaining valid half-body combinations in ascending order of the approximate symmetric chamfer distances between the combinations and the input contour. The top l ranked combinations are marked as candidate combinations.

Fine Evaluation of Candidate Combinations Re-sort the candidate combinations based on fine matching for contour, silhouette, and structural cues. Given this final ranking, we select the highest or a few top ranked combination(s) as the pose estimate.

3 Half-body Pose Retrieval

We use chamfer distance [1] to retrieve the half-body poses for a given query image. Chamfer distance is an efficient tool to measure dissimilarity between contours because it does not explicitly make point corre-

spondences between contours and yet is robust again minor misalignment in position and scale. Moreover, chamfer distance can allow matching partial to whole contours. We use this advantage to retrieve half-body candidate poses from the database.

3.1 Half-body Pose Database

The database consists of a large number of half-body poses and corresponding CG renderings. The database was created from a set of motion capture data taken from a public website [4]. From a total of 13,000 motion data frames, 1247 key frames were evenly selected including walking, running, playing basketball, and dancing. With a famous animation and 3D character design package called Poser [2], for each selected key frame, upper-body and lower-body poses were respectively rendered from 12 equal-spaced virtual cameras, yielding 14,964 upper-body and 14,964 lower-body human silhouettes. Subsequently, contours and DTs were derived from silhouettes. All of these data — including silhouettes, contours, and DTs, plus 3D poses — were stored in the database. For space efficiency, silhouettes and contours were stored in an efficient format, recording only the location information of the foreground pixels. The upper-body pose was composed of 3D locations for 10 upper-body joints and the lower-body pose for the remaining 7 lower-body joints.

3.2 Chamfer Distance

We denote query contour $Q = \{q\}$ and the half-body contour in database $T = \{t\}$, both represented by a set of points on the boundary. Chamfer distance $d_{cham}^{(T,Q)}$ is calculated by taking the mean distance of all points in T to their closest points in Q:

$$d_{cham}^{(T,Q)} = \frac{1}{N_T} \sum_{t \in T} \min_{q \in Q} ||t - q||, \quad (1)$$

where N_T is the number of points in T and $|| \cdot ||$ can be any norm, e.g., Euclidian, Cityblock, etc. Chamfer distance $d_{cham}^{(T,Q)}$ can be efficiently computed by first evaluating the DT of contour Q using a two-pass algorithm proposed in [1]. The value of each pixel in the DT equals the closest distance from that pixel to contour Q. τ-DT, which truncates a large distance by threshold value τ, is more frequently used in practice. By substituting τ-DT of contour Q,

$$DT_Q^\tau(t) = \min(\min_{q \in Q} ||t - q||, \tau), \quad (2)$$

into (1), chamfer distance has new form

$$d_{cham}^{(T,Q)} = \frac{1}{N_T} \sum_{t \in T} DT_Q^\tau(t), \quad (3)$$

where a simple lookup operation replaces the time-consuming *min* operation in (1).

3.3 Retrieving Half-body Candidates

Chamfer distances are computed from all half-body examples in the database to the query contour. We sort the database examples in ascending order of computed distances, and then select the top k ranked examples as candidates for upper- and lower-bodies.

However, choosing irrelevant half-body poses for the simple but imperfect chamfer distance is likely. For instance, some upper-body contours might happen to match the lower-body part of the query contour. Consequently, many irrelevant half-body poses may be mistakenly chosen as candidates. We handle this problem by considering a sufficient number of candidates. In current retrieval step, only sufficient relevant half-body candidates must be found, regardless of how many irrelevant ones are included. Subsequent steps can prune the irrelevant half-body candidates using further matching. We found that retaining several hundred half-body poses as candidates is usually sufficient.

4 Pruning Invalid Combinations

After retrieving half-body candidates, we determine the optimal half-body combination(s) generated by these candidates. However, exhaustive searching

through hundreds of thousands of possible combinations requires intensive computation. Fortunately, most combinations are invalid 3D human poses and can be efficiently ruled out under a combination constraints learned from database examples.

In this work the combination constraints are represented simply as a 2D table C: the value of each entry C_{ij} is binary indicating whether the combination of upper-body pose i and lower-body pose j forms a valid full-body pose. Based on deviation of body orientation and similarity of pose, we determine the combinative lower-body poses for every upper-body pose in the database. The algorithm details are omitted due to the space limitation.

Combination constraint allows about 10 million valid combinations. Under this constraint, in general $10,000 \sim 30,000$ valid combinations remain from hundreds of thousands of exhaustive (unconstraint) combinations.

5 Coarse Evaluation

This step aims to coarsely evaluate valid half-body combinations by which we choose a set of candidate combinations. Coarse evaluation is based on the symmetric chamfer distance between the query contour and the contour of combination. The symmetric chamfer distance is just the sum of bi-directional chamfer distances, defined as

$$D_{cham}^{(UL,Q)} = d_{cham}^{(UL,Q)} + d_{cham}^{(Q,UL)}, \quad (4)$$

where UL denotes the contour of half-body combinations. Symmetric chamfer distance generally outperforms the (directed) chamfer distance used in retrieval steps.

Computing the exact symmetric chamfer distance of $10,000 \sim 30,000$ combinations creates an issue of high computational cost because it involves a set of time-consuming operations, including half-body silhouette combination, contour extraction, and DT calculation. However, we can resort to an approximate distance that can avoid such time-consuming operations. The idea is simple. We approximate the contour of combination as the union of half-body contours, that is, $UL \simeq U \cup L$. When upper-body and lower-body figures are completely separated, the contour of combination equals the union set of half-body contours. Although more frequently equalization does not hold true due to half-body overlapping, the approximation is reasonable because overlapping is not severe in most cases.

Consequently, chamfer distance $d_{cham}^{(UL,Q)}$ is approximated as

$$d_{cham}^{(UL,Q)} \simeq \frac{1}{N^U + N^L}\left(\sum_{u \in U} DT_Q^r(u) + \sum_{l \in L} DT_Q^r(l)\right), \quad (5)$$

where $\sum_{u \in U} DT_Q^r(u)$ and $\sum_{l \in L} DT_Q^r(l)$ are two quantities already calculated during the retrieval step.

On the other side, chamfer distance $d_{cham}^{(Q,UL)}$ is approximated as

$$d_{cham}^{(Q,UL)} \simeq \frac{1}{N^Q} \sum_{q \in Q} \min(DT_U^r(q), DT_L^r(q)), \quad (6)$$

which involves several hundred min operations.

Apparently, the above approximate symmetric chamfer distance is very efficient to compute. We rank valid half-body combinations in ascending order of computed approximate distances. Then a set of top ranked combinations is chosen for further fine evaluation.

6 Fine Evaluation

Given the short list of remaining candidate combinations, we want to re-rank them based on fine matching, which involves three types of distances: edge, silhouette, and structural cues.

Edge Distance Edge distance is the approximate symmetric chamfer distance plus an additional correspondence penalty. Since chamfer distance does not ensure uniform-point correspondence between two contours, we introduce a penalty to punish unbalanced correspondence, defined as

$$p_{corres}^{(UL,Q)} = 1 - \frac{N^{[C]_u}}{N^{UL}}, C = \{c | \operatorname*{argmin}_{c \in Q} ||c - x||, \forall x \in UL\} \quad (7)$$

where $N^{[C]_u}$ is the number of unique points in the set of corresponding points C and N^{UL} is the number of points on combined contour UL. Set C can be obtained effortlessly because the correspondence information for Q has been already stored during the calculation of DT. The $p_{corres}^{(Q,UL)}$ can be obtained likewise, but it is not used for computational efficiency.

Silhouette Distance The silhouette of half-body combination is computed by pixel-wise logical OR operation on two half-body silhouettes. Silhouette distance is proportional to the nonoverlapping area, defined as

$$D_{silhouette}^{(Q,UL)} = 1 - \frac{AREA(S^{UL} \bigwedge S^{Q})}{\max(AREA(S^{Q}), AREA(S^{UL}))}, \quad (8)$$

where S^{UL} and S^{Q} denote the silhouettes of half-body combination and query, respectively.

Structural Distance We also use Hu moment to characterize the structural information of silhouettes. The distance is defined as

$$D_{structure}^{(Q,UL)} = \sqrt{\frac{\sum_{i=1}^{10}(h_i^{Q} - h_i^{UL})^2}{10}}, \quad (9)$$

where h^{Q} and h^{UL} are 10 Hu moments derived from silhouettes S^{Q} and S^{UL}, respectively.

The overall distance is the weighted average of the above three distances:

$$D_{cham}^{(Q,UL)} + \lambda p_{corres}^{(UL,Q)} + \beta D_{silhouette}^{(Q,UL)} + \gamma D_{structure}^{(Q,UL)}, \quad (10)$$

where weight coefficients λ, β, and γ are set appropriately as 0.5, 1.5, and 2.0 through cross validation.

We re-rank half-body combinations in ascending order of overall distance. From the ranked results, we choose the highest or top ranked combinations as the pose estimate.

7 Experiments

To evaluate our method's performance, we conducted experiments using both synthetic and real images. The synthetic image experiment provided a quantitative evaluation to our method because 3D ground truth poses are known. The real image experiment shows the generalization ability of our method because a wide range of 3D poses are included, and variations in clothing, body size, and view angle complicate the estimation task. In the experiments, 800 high ranked upper-body and 300 high ranked lower-body poses were selected in the retrieval step and fed into the coarse evaluation step. Next 100 high ranked half-body combinations were selected to go through fine evaluation.

7.1 Experiment with Synthetic Images

We quantitatively evaluated our method using k-fold cross validation on synthetic images. The dataset of synthetic images (created from identical motion data with a half-body database) is divided into 7 subsets (around 2100 samples in each subset). In each trial, one of the 7 subsets is used as the test set, and corresponding half-body data are excluded from the database during retrieval. Then average error over all 7 trials is computed.

The root mean square error (RMSE) of all 3D joint positions (divided by the number of joints) is calculated between the ground truth and the selected combination of half-body poses. To better analyze the resulting errors, we calculated reference error by taking the mean of RMSE errors of successive pose pairs in the database. Since the database is created from motion data, successive poses usually resemble each other. We chose 95% from all pairs with lower error to exclude outliers.

We evaluated performance in two manners: rank-1 (highest rank) combination, and the best of top 5 ranked combinations (based on ground truth). Fig.1 shows average errors obtained with/without fine evaluations for full body, upper/lower body, and parts. Clearly the errors incurred by the top 5 are generally lower than reference error. Table 1 presents the ratio of good estimation for each case, as in Fig.1. Here, pose estimation is judged good if error between it and ground truth is lower than reference error. It can be concluded from the data in Table 1 that fine evaluation generally gains 2 points of performance improvement over coarse evaluation; the

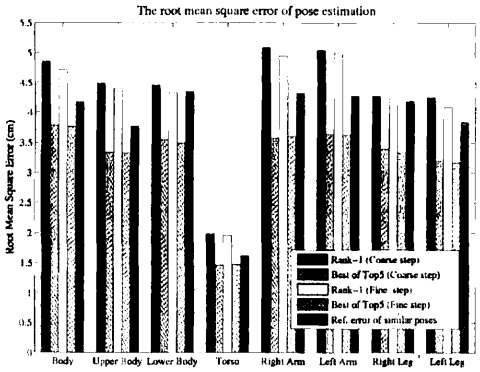

図 1: Performance evaluation using cross validation

表 1: Ratio of good pose estimation

Method	Coarse Evaluation		Fine Evaluation	
	Rank-1	Top5	Rank-1	Top5
Full Body	55%	67%	57%	69%
Upper Body	59%	70%	60%	72%
Lower Body	67%	75%	67%	77%
Torso	63%	75%	65%	76%
Right Arm	59%	73%	61%	74%
Left Arm	62%	74%	63%	75%
Right Leg	68%	76%	68%	77%
Left Leg	66%	76%	66%	76%

best of top5 significantly outperformed rank-1 by 11 points on average. The achieved ratios of good estimation are less than 80%, which does not mean poor performance because the chosen reference error is a very strict value. We conducted a random selection experiment by randomly choosing 5 poses from the database and then selecting the minimum error based on ground truth. We achieved 25 good estimations out of 10,000 trials. In other words, the ratio of good estimation by random selection is 0.25%.

7.2 Experiment with Real Images

We also conducted an experiment on a set of 155 real images collected using Google's image search engine. The test images involve a variety of sports: basketball, dance, football, tennis, figure skating, etc. We first used graphic software to semi-automatically extract silhouettes. Many of these images were difficult (even for people) to infer the underlying 3D human poses from only silhouettes. The segmented images were automatically normalized using a regression-based method (that is not described here due to the space limitation). To avoid the problems of inverse body orientation, we assume that body orientation is known to be outward or inward of the image plane.

Fig.2 shows examples of estimation results obtained by our method. For each test image, 3D rendering of the rank-1 combination is shown, and its view rotated 90 degree is also shown to clearly observe 3D rendering. Notice that, even though a large amount of information for describing human body poses was lost in normalized silhouettes, the results are visually close to what can be considered the right pose for input images. Good estimations are commonly achieved for lower-body poses, but biased estimates of upper-body poses sometimes result from high occlusion. See such examples as Figs.2(b), 2(d), and 2(e) in the silhouettes where arms are strongly occluded and precisely estimating upper-body pose is quite difficult. However, when we consider more combinations than only the rank-1 combination, clear improvements are achieved. Fig.3 shows such example. In the figure, the last three boxes in turn show rank-1, rank-2, and rank-3 combinations. We found that the rank-3 combination is clearly much closer to the right pose than the rank-1 combination.

Since there are no ground truth poses for these real images, we can only subjectively evaluate the quality of resulting estimations. We built a browser-based rating system that presents the top 3 ranked combinations for each test image to participants who subjectively find the best of three combinations and rate the quality with four levels: *great* (4 points) when the estimate is en-

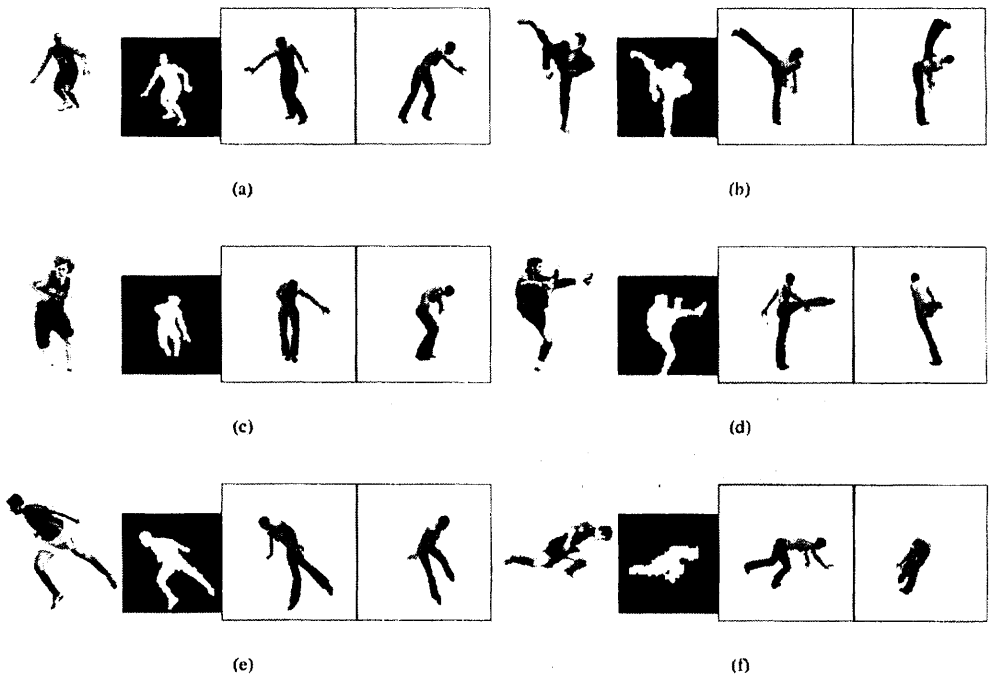

図 2: Examples of (Rank-1 based) 3D pose estimation on real images. In each set of images, the first two are the original image and the normalized silhouette, and the last two are the CG rendering of the estimated pose and the view rotated 90 degree, respectively.

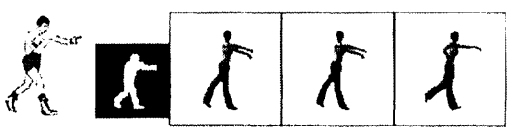

図 3: An example showing better estimation in latter ranked combinations

tirely consistent to human perception, *good* (3 points) when one half-body estimation is good but the other is less biased, and *average* (2 points) when one half-body estimation is good but the other is strongly biased, and *bad* (1 point) when no good half-body estimation is found. A summary of the subjective evaluations are shown in Fig.4. The average rate over all test images is 2.7 points. The evaluation for some categories such as figure skating is good, although unrelated to the action categories used in the database. However, some other categories such as kung fu and golf are not so good due to indistinct contour (caused by image downsampling or occlusion) and very complex poses. Perhaps this performance is promising, considering the complexity of the task and the simple image information used.

7.3 Time and Space Complexities

The method was mainly implemented using Matlab 7. Some algorithms such as half-body retrieval and coarse evaluation are implemented by C++. The PC for running the experiments had an Intel Pentium 4 CPU running at 3.2 GHz and 1 GB RAM. Computational cost for

表 2: Empirical computational complexity and memory requirement

Image Proc.	Retrieval	Coarse Eval.	Fine Eval.	DTs	Constraints	Silhouettes	Contours
0.05s	0.14s	0.45s	0.34s	285MB	213MB	34MB	39MB

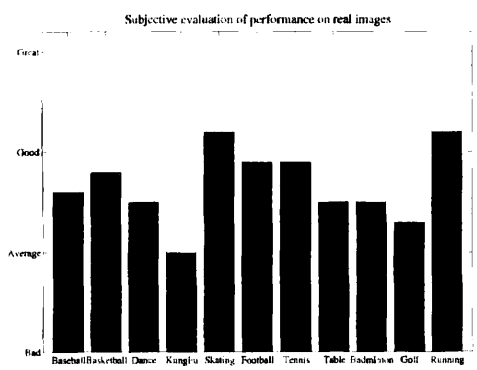

図 4: Subjective evaluation of performance on real images. From left to right: Baseball (14), Basketball (31), Dance (11), Kung Fu (18), Figure Skating (13), Football (32), Tennis (16), Table Tennis (4), Badminton (6), Golf (4), and Running (6). The number in parenthesis indicates data size.

the overall process was approximately 1 second per test image. Around 600 MB of memory are required to run this system. Table 2 shows the details of computational time and memory requirement.

8 Conclusion and Future Work

We presented a retrieval-combination approach for estimating 3D human pose from a monocular image. Although only image observations of human silhouettes are used, our approach produces satisfying results for a wide variety of human poses. The basic strategy of our approach retrieves half-body candidate poses for a given image by partial contour matching. Then under learned combination constraint, invalid combinations of half-body candidates are ruled out, and from the remaining valid combinations we efficiently choose the most likely one or a short list by coarse-to-fine evaluation.

Augmenting the database with many more 3D human poses and exploring the use of other informative observations, such as skin color or face-like features, will further improve performance. Handling a wider range of complex poses while maintaining computational efficiency is the main direction of our future research. In addition, we are interested in applying our approach to clutter images.

参考文献

[1] H.G. Barrow, J.M. Tenenbaum, R.C. Bolles, and H.C. Wolf. Parametric correspondence and chamfer matching:two new techniques for image matching. *IJCAI*, pages 659–663, 1977.

[2] e frontier / Curious Labs. Poser 6: The premiere 3d figure design and animation solution, 2005.

[3] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1), 2005.

[4] BVH File Repository. http://www.centralsource.com/blender/bvh/.

[5] R. Rosales and S. Sclaroff. Specialized mappings and the estimation of human body pose from a single image. In *IEEE Workshop on Human Motion*, pages 19–24, 2000.

[6] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. In *ICCV*, volume 2, pages 750–757, 2003.