Pictorial Structures on RGB-D Images for Human Pose Estimation in the Operating Room

Abdolrahim Kadkhodamohammadi[1], Afshin Gangi[1,2],
Michel de Mathelin[1], and Nicolas Padoy[1]

[1] ICube, University of Strasbourg, CNRS, IHU Strasbourg, France
{kadkhodamohammad,gangi,demathelin,npadoy}@unistra.fr
[2] Radiology Department, University Hospital of Strasbourg, France

Abstract. Human pose estimation in the operating room (OR) can benefit many applications, such as surgical activity recognition, radiation exposure monitoring and performance assessment. However, the OR is a very challenging environment for computer vision systems due to limited camera positioning possibilities, severe illumination changes and similar colors of clothes and equipments. This paper tackles the problem of human pose estimation in the OR using RGB-D images, hypothesizing that the combination of depth and color information will improve the pose estimation results in such a difficult environment. We propose an approach based on pictorial structures that makes use of both channels of the RGB-D camera and also introduce a new feature descriptor for depth images, called histogram of depth differences (HDD), that captures local depth level changes. To quantitatively evaluate the proposed approach, we generate a novel dataset by manually annotating images recorded from different camera views during several days of live surgeries. Our experiments show that the pictorial structures (PS) approach applied on depth images using HDD outperforms the state-of-the art PS approach applied on the corresponding color images by over 11%. Furthermore, the proposed HDD descriptor has superior performance when compared to two other classical descriptors applied on depth images. Finally, the appearance models generated from the depth images perform better than those generated from the color images, and the combination of both improves the overall results. We therefore conclude that it is highly beneficial to use depth information in the pictorial structure model and also for human pose estimation in operating rooms.

Keywords: Body pose estimation, Pictorial structures, RGB-D images, Surgical workflow analysis.

1 Introduction

Recognizing and estimating the poses of clinical staff can provide invaluable input for many applications in the operating room (OR), for example surgical

N. Navab et al. (Eds.): MICCAI 2015, Part I, LNCS 9349, pp. 363–370, 2015.
DOI: 10.1007/978-3-319-24553-9_45

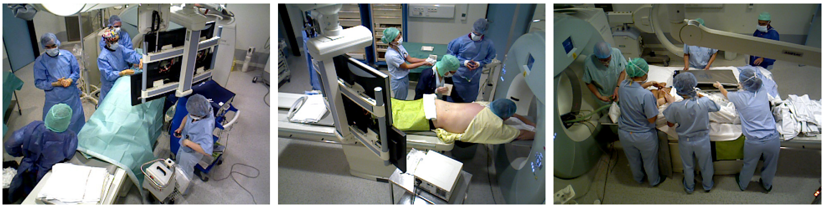

Fig. 1. Sample images of the clinician pose dataset acquired during live surgeries. Each image shows the OR from one of three different views used to capture the environment.

activity recognition [8,4], performance assessment [6], context-aware user interfaces [9,7] and radiation safety monitoring [5].

Human pose estimation (HPE) is also one of the fundamental problems of computer vision and has produced a vast literature over the last few decades. HPE is mainly addressed by part based approaches thanks to [2] that has made exact inference tractable on pictorial structures (PS) with tree-structured graphs by using the generalized distance transform algorithm. Recently, Yang and Ramanan [13] proposed an approach based on structured support vector machines (SVM) and PS that jointly learns mixtures of parts and a loosely coupled pairwise dependency model to capture different part appearances and their co-occurrences. Other works have tried to improve PS by enhancing its underlying graphical model to capture higher order relationships between parts, such as temporal consistency [3] and repulsive factors between left and right parts [1], with the penalty of approximate inference. Toshev at al. [12] also proposed a holistic method based on deep neural networks to regress for body joints, which requires a very large training set. In contrast to previous works that are only relying on color images to model body part appearances, [10] proposed to use random forests on depth images alone and reported promising result for human pose estimation in scenarios where a foreground mask can be obtained. However, obtaining foreground masks in the cluttered scenes of the OR, as shown in figure 1, is not a trivial task.

Even though many HPE methods are proposed in the vision literature, they cannot be directly applied to the OR due to the numerous difficulties implied by such an environment. First of all, the safety requirements of the OR and the multiple articulated arms mounted on the ceiling constrain camera positioning. Second, the illumination can change severely. Finally, most equipments and clinical clothes have the same color. [3] proposes an approach for 3D human pose estimation in the OR using discrete optimization over RGB-D sequences. A generic body part detector was used, similar to [10], that had not been trained for the OR environment. To cope with the misdetections of the detector and to enforce temporal consistency, [3] proposes an optimization over a connected graph modeling the whole sequence. Ground-truth initializations for the first and last frame are needed. These requirements make the approach unsuitable for real-time applications. Moreover, the dataset used for evaluation has been recorded in real interventional rooms but not during real surgeries. Thus, it does not include all the visual challenges occurring during live operations.

In this work, we address the challenging problem of human pose estimation in the OR. We base our approach upon [13] due to four main reasons: first, it is the state-of-the-art method for human pose estimation; second, by using mixtures of parts, the approach tolerates appearance changes and part foreshortening, which are common in the OR; third, the method performs exact inference in a tractable manner on tree-structured pairwise dependency models; and finally, all parameters are learned using a structured SVM solver, thus eliminating the need for further manual adjustments.

The body part appearance models in [13] are built using intensity images only. The histogram of oriented gradient (HOG) descriptor, referred to as I-HOG in this paper when applied on intensity images, is used to build the part models. However, in a complex environment such as the OR, color images might not always carry enough descriptive information due to severe illumination changes and high color similarities. With the advent of affordable RGB-D sensors, it is now possible to capture both color and depth images simultaneously. The depth image is computed by decoding the deformation of a known pattern that is projected onto a scene using infrared light. Thus, the depth accuracy is not affected by brightness or contrast changes. We therefore propose to construct robust and discriminative body part appearance models by combining both color and depth features.

We also introduce a new descriptor for depth images, named histogram of depth differences (HDD), that encodes surface level changes. We compare it with two other descriptors for depth images: the aforementioned HOG descriptor that we apply on depth images for comparison (D-HOG), and the histogram of oriented normal vectors (HONV) [11] that has been reported to perform well for object detection in depth images.

In summary, we propose an approach for human body pose estimation in the OR that relies jointly on color and on depth images. To the best of our knowledge, this is the first PS-based approach on RGB-D data and also the first single frame human pose estimation approach that has been proposed for the OR. We additionally introduce the HDD descriptor for depth images and compare it with two other classical descriptors. Finally, we quantitatively evaluate our approach on a novel dataset recorded during several days of live surgeries.

2 Method

In this section, we briefly describe the flexible mixtures of parts (FMP) approach [13] that serves as a basis for our human pose estimator. We also present the HONV descriptor and introduce the HDD descriptor for depth images.

2.1 Flexible Mixtures of Parts

The flexible mixtures of parts approach [13] represents the human body as a set of rigid body parts that are loosely coupled with each other using springs. The state of each body part i is specified by the pixel position l_i of its joint and by its

0	0	0
-1	0	1
0	0	0

-1	0	0
0	0	0
0	0	1

0	-1	0
0	0	0
0	1	0

0	0	-1
0	0	0
1	0	0

Fig. 2. Four different kernels that capture local level changes in depth images.

body part type t_i. This type can be defined based on low level information, such as part position, or high level information, such as the state of the part (e.g. open versus closed hand). We write $i \in \{1, ..., K\}, l_i \in \{1, ..., L\}$ and $t_i \in \{1, ..., T\}$. Let $G = (V, E)$ be a relational graph whose K nodes represent body parts and whose edges define connections between body parts to enforce body kinematics.

$$S(I, D, l, t) = \sum_{i \in V} w_i^{t_i}.\phi(I, D, l_i) + \sum_{ij \in E} w_{ij}^{t_i,t_j}.\psi(l_i - l_j) + \sum_{i \in V} b_i^{t_i} + \sum_{ij \in E} b_{ij}^{t_i,t_j} \quad (1)$$

is a score function, where I denotes the input image and D is the aligned depth image. The first term in eq. 1 is the part appearance model that computes the score of placing a part i at location l_i based on the part appearance template $w_i^{t_i}$ and the feature vector $\phi(I, D, l_i)$. The second term computes the deformation score based on relative displacements between parts that are encoded by $\psi(l_i - l_j) = [dx\ dx^2\ dy\ dy^2]^T$, where (dx, dy) is the relative location of part i w.r.t. part j. The last two terms are computing the part type compatibility score, where b_i is the score of choosing a particular mixture for part i and $b_{ij}^{t_i,t_j}$ encodes the co-occurrence compatibility of part types.

Inference and Learning: Estimating the body pose in this model corresponds to finding $(l^*, t^*) = \text{argmax}_{l,t}\, S(I, l, t)$. For a tree-structured relational graph G, this optimization can be solved efficiently and exactly with dynamic programming. In a supervised learning setup, we assume that a set of labeled positive and negative examples is available. The model parameters can then be learned using structural SVM. For more details, we refer to the original paper [13].

2.2 Appearance Model

We now focus on the different feature descriptors that can be used to build the body part appearance model. To construct ϕ, we uniformly divide images into non overlapping windows, called cells, in which the descriptors are computed. Following the same avenue as [13], I-HOG is used on color images and serves as our baseline. The D-HOG descriptor applies HOG on depth images. The two other depth-based descriptors used in this paper are described below.

Histogram of Oriented Normal Vectors (HONV): This descriptor represents object surfaces using a histogram of normal vectors. The gradient of the depth image is first used to compute the normal vectors in spherical coordinates. Each normal vector is then represented by a pair of azimuth and zenith angles and quantized to vote into a 2D histogram; more details can be found in [11].

Histogram of Depth Differences (HDD): A depth image encodes in each pixel the distance between the depth sensor and the surfaces of the scene. In [10], a simple depth operator is proposed that performs well for body part detection. This operator is able to distinguish between different body parts by capturing part depth differences and is learned during the training of a deep random forest. Inspired by this operator, we introduce a simple yet efficient new descriptor that captures local level changes in a depth image using the kernels shown in figure 2. Let K_k be one of the HDD kernels with $k \in \{1, ..., 4\}$. We define the normalized convolution response at position (x, y) as

$$C_{ks}(x, y) = (K_k \, * \, P_s(x, y))/D_s(x, y), \tag{2}$$

where $s \in \{1, ..., n\}$ is the scale of the depth image and $P_s(x, y)$ is the depth image patch at location (x, y) for scale s. The convolution is applied over a scale space to capture the changes in different spatial neighborhoods. The responses are also normalized by the depth value at the center of each patch for depth invariance. For each cell, the descriptor is then built by quantizing all the responses and by binning them into a 3D histogram of kernel, scale and quantization levels.

3 Experimental Results and Discussions

We evaluate our approach on a new RGB-D dataset recorded in an operating room using an *Asus Xtion Pro Live* camera. To capture different views of the room, the camera position is changed among three possible locations, as shown in figure 1. All activities happening in the OR are recorded for seven half days at 20fps. Due to the large number of frames and the similarity of consecutive frames, we manually annotate every 500th frame to provide ground truth positions for the upper body parts. Ultimately, we obtain a dataset that includes 1451 frames containing 1991 annotated persons and that is balanced across the half days. We have also selected 173 frames that do not contain any human to serve as negative examples. Since the distances of the annotated persons with respect to camera vary in the range of $[1.2 - 4.5]$ meters, body parts at multiple scales are present in this dataset. Different individuals including surgeons, nurses and anesthesiologists wearing various clothes, such as scrub suits, surgical gowns and radiation protection aprons have been annotated. We have divided the dataset into seven disjoint sets where all frames in a set belong to the same half day recording. A leave-one-out scheme is applied so that one set is used as test set and the rest as training set. We report the average results of seven-fold cross validation during the evaluation.

Following current practice in the literature, we use the probability of correct keypoint (PCK) as evaluation metric [13]. A tight bounding box is obtained from the annotated upper body pose of each person. Given the bounding box, the pose estimator returns a set of keypoints indicating the body joint locations. A keypoint is correct if it falls within $\alpha \, . \, max(w, h)$ pixels of the ground truth, where w and h are respectively the width and height of the bounding box, and α specifies the relative threshold used to consider a keypoint as a correct detection.

Table 1. PCK results for different appearance models on the clinician pose dataset.

Color Desc.	Depth Desc.	Head	Shoulder	Elbow	Wrist	Hip	Total
I-HOG	-	84.12	72.70	56.98	56.53	45.91	63.25
-	D-HOG	90.87	83.30	65.15	66.21	57.12	72.53
I-HOG	D-HOG	**93.92**	84.75	70.26	67.80	59.54	75.25
-	HONV	84.80	80.20	58.64	56.81	47.52	65.60
I-HOG	HONV	92.52	**85.89**	69.31	68.91	60.27	75.38
-	HDD	91.18	82.89	67.35	69.93	62.12	74.69
I-HOG	HDD	92.76	84.10	**71.12**	**71.62**	**63.62**	**76.64**

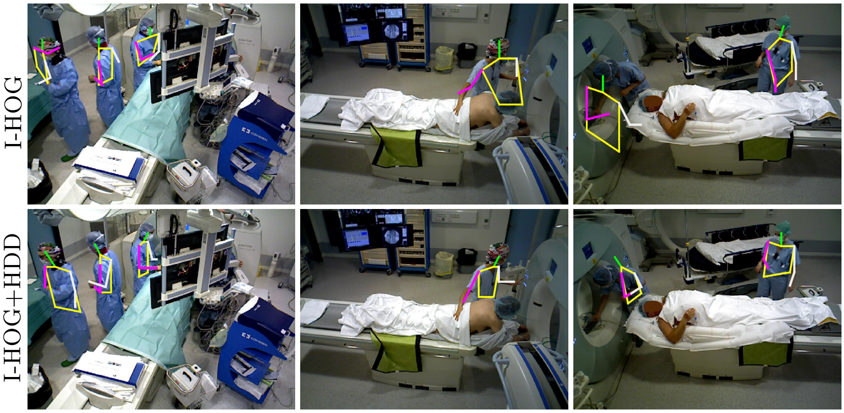

Fig. 3. Examples of pose estimation using I-HOG and I-HOG+HDD.

We set $\alpha = 0.2$ as proposed by [13]. In table 1, we present the evaluation results for the different appearance models on the clinician pose dataset. To combine different appearance models, we concatenate their descriptors. A cell size of 6×6 pixels has been set to stay consistent with the one commonly used with HOG.

HOG: We construct the body part appearance models using the HOG descriptor in three different ways. We first evaluate HOG on color images (I-HOG) to generate the baseline using the same parameters as in [13]. Second, we apply the same descriptor on the depth images and follow the same pipeline (D-HOG). Finally, both I-HOG and D-HOG are used to jointly learn body part models. D-HOG improves the results over the baseline significantly (+9.28%), which indicates the strength of depth-based appearance models as compared to the original color-based one. Since the same descriptor is used for both I-HOG and D-HOG, it highlights that the intensity gradient is not always reliable in such an environment due to the high color similarity of the clothes and equipments as well as the severe illumination changes, while depth gradients provide more robust representations. The representation based on I-HOG+D-HOG boosts the performance even more (+12%) by building a strong appearance model that combines the complementary information provided by the color and depth sensors.

HONV: To evaluate the performance of HONV, a normalization scheme is needed for the descriptor. We have compared the L2-norm and L2-Hys normalization schemes, where L2-Hys is similar to L2-norm but limits the maximum value to 0.2. Since both normalization methods show similar performance and L2-Hys is used with HOG, only the results with the L2-Hys normalization scheme are reported. HONV improves the results (+2.35%) over the baseline and its combination with I-HOG enhances the performance (+12.13%).

HDD: We compute the HDD descriptor in three scales and coarsely quantize the kernel responses into 10 levels to be robust to noise and spatial distortions. Thus, we obtain a descriptor of size 4 × 3 × 10 per cell. Since L2-norm and L2-Hys normalizations show similar performance, following the same reasoning as before, we report the results with L2-Hys normalization. As shown in table 1, the HDD descriptor significantly improves the results over the baseline (+11.44%). Combining both HDD and I-HOG further boosts the performance (+13.39%).

In summary, the experiments show that the color descriptor I-HOG is not always reliable due to the lack of texture on the surgical clothes and to the color similarities of clothes and equipments, while the depth descriptors provide more robust body part representations. Table 1 also shows the results per body part. High improvements on shoulder detection are obtained using all depth descriptors. This can be explained by the very distinctive depth changes on shoulders that are often not captured by the color images due to color similarities or illumination changes. However, noisy depth maps can heavily distort edges and surface normals, especially when body parts are close to each other. The consistent performance improvement of HDD over the other depth descriptors on the most challenging parts, namely elbow and wrist, shows the benefits of its coarse representation in such cases. Finally, the combination of color and depth descriptors improves the body part models and the best performance is obtained using I-HOG+HDD. Figure 3 shows the overlaid skeletons for several frames using I-HOG and I-HOG+HDD. The first two columns illustrate cases where the color-based body part detector is confused by intensity changes while the HDD-based detector correctly localizes the body parts. The last column shows failure cases for both descriptors that indicate the need for better occlusion handling and 3D inference.

4 Conclusions

In this paper, we propose an approach based on pictorial structures for human pose estimation in the operating room. We first extend the appearance representations used in PS to RGB-D images by constructing strong and discriminative body part detectors using both color and depth images to deal with the visual challenges present in the room. We also propose a new descriptor for depth images that encodes the local level changes in a 3D histogram. Since this descriptor is computed over a scale space and coarsely discretized, it is robust to local geometric distortions. Furthermore, to quantitatively evaluate our approach, we generate a novel RGB-D dataset in which upper body poses of clinicians are manually annotated. We then conduct a series of experiments that compare the different appearance representation approaches. We show that depth descriptors are performing better than the color descriptor. Furthermore, the combination of the color and depth descriptors always improves the body part detections. Finally, the proposed HDD descriptor improves the performance by 13.39% over the baseline and significantly enhances the elbow, wrist and hip detections.

Acknowledgements. This work was supported by French state funds managed by the ANR within the Investissements d'Avenir program under references ANR-11-LABX-0004 (Labex CAMI), ANR-10-IDEX-0002-02 (IdEx Unistra) and ANR-10-IAHU-02 (IHU Strasbourg).

References

1. Andriluka, M., Roth, S., Schiele, B.: Discriminative appearance models for pictorial structures. Int. J. Comput. Vision 99(3) (2012)
2. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. Int. J. Comput. Vision 61(1), 55–79 (2005)
3. Kadkhodamohammadi, A., Gangi, A., de Mathelin, M., Padoy, N.: Temporally consistent 3D pose estimation in the interventional room using discrete MRF optimization over RGBD sequences. In: Stoyanov, D., Collins, D.L., Sakuma, I., Abolmaesumi, P., Jannin, P. (eds.) IPCAI 2014. LNCS, vol. 8498, pp. 168–177. Springer, Heidelberg (2014)
4. Lea, C., Facker, J.C., Hager, G.D., Taylor, R.H., Saria, S.: 3D sensing algorithms towards building an intelligent intensive care unit. In: AMIA CRI (2013)
5. Loy Rodas, N., Padoy, N.: Seeing is believing: increasing intraoperative awareness to scattered radiation in interventional procedures by combining augmented reality, monte carlo simulations and wireless dosimeters. International Journal of Computer Assisted Radiology and Surgery, 1–11 (2015)
6. Mason, J., Ansell, J., Warren, N., Torkington, J.: Is motion analysis a valid tool for assessing laparoscopic skill? Surgical Endoscopy 27(5), 1468–1477 (2013)
7. Noonan, D.P., Mylonas, G.P., Darzi, A., Yang, G.Z.: Gaze contingent articulated robot control for robot assisted min. invasive surgery. In: IROS (2008)
8. Padoy, N., Mateus, D., Weinland, D., Berger, M.O., Navab, N.: Workflow Monitoring based on 3D Motion Features. In: VOEC-ICCV, pp. 585–592 (2009)
9. Schwarz, L., Bigdelou, A., Navab, N.: Learning gestures for customizable human-computer interaction in the operating room. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part I. LNCS, vol. 6891, pp. 129–136. Springer, Heidelberg (2011)
10. Shotton, J., Fitzgibbon, A.W., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR, pp. 1297–1304 (2011)
11. Tang, S., Wang, X., Lv, X., Han, T., Keller, J., He, Z., Skubic, M., Lao, S.: Histogram of oriented normal vectors for object recognition with a depth sensor. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part II. LNCS, vol. 7725, pp. 525–538. Springer, Heidelberg (2013)
12. Toshev, A., Szegedy, C.: DeepPose: Human pose estimation via deep neural networks. In: CVPR (2014)
13. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. PAMI 35(12), 2878–2890 (2013)