OSiMa: Human Pose Estimation from a Single Image

Nipun Pande and Prithwijit Guha

TCS Innovation Labs, New Delhi
{nipun.pande,prithwijit.guha}@tcs.com

Abstract. Human upper body pose estimation plays a key role in applications related to human-computer interactions. We propose to develop an avatar based video conferencing system where a user's avatar is animated following his/her gestures. Tracking gestures calls for human pose estimation through image based measurements. Our work is motivated by the pictorial structures approach and we use a 2D model as a collection of rectangular body parts. Stochastic search iterations are used to estimate the angles between these body parts through Orientation Similarity Maximization (OSiMa) along the outline of the body model. The proposed approach is validated on human upper body images with varying levels of background clutter and has shown (near) accurate pose estimation results in real time.

1 Introduction

Human pose estimation through visual sensing has potential applications in the fields of human computer interfaces and motion capture for realistic 3D animations, biomechanical analysis, robot control and visual surveillance (activity recognition). Our work is aimed at developing an avatar based video conferencing system where a user's avatar in a virtual meeting room is animated following his/her gestures. Here, we present a part of this work where we restrict ourselves to the pose estimation of *human upper body* only while assuming a single person in the view. Tracking human poses through image based measurements has two major sub-divisions – first, the (near) accurate estimation of the human pose from the very first image and second, re-estimation of the pose from the subsequent images using the previous instant's pose as a prior. Here, we focus on the necessary first step, i.e. human upper body pose estimation from a single image to initialize the gesture tracking procedure.

Existing literature on human body pose estimation is vast [1], most of which assume the existence of a background model to handle the scene clutter. However, we restrict our discussions to approaches aimed at retrieving human upper body pose from single images only (i.e. no temporal information). Ramanan describes an iterative parsing process to estimate the pose using region based body models e.g. color histogram to refine the body part positions [2]. An approach following this is [3]. Here an upper body detector is used with GrabCut to determine the foreground area followed by parsing to fit a pictorial structure

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 200–205, 2011.

model. An interesting work which takes into account natural shape and pose variations is Contour People [4]. Here segmentation of the scene into foreground and background regions is done and a shape deformation model is trained using 3D SCAPE model.[5] express a subset of model parameters as kernel regression estimates from a learned sparse set of exemplars. An exemplar based pose representation called "poselet" is used in [6]. To detect the presence of each poselet a classifier is trained for each poselet using standard linear SVM and the histogram of oriented gradients.

We have adopted a human body model consisting of rectangular body parts (head, torso and lower/upper arms) with circles at joints to render a smooth body contour in the image plane. A Haar feature based frontal face detector [7] is used to localize the head in the image. The dimensions of the different body parts are computed from the head (or face) width using available anthropometric ratio data [8]. Now, considering the head position as pivot, several body model outlines can be rendered by varying the angles between the body parts. This provides us with a set of straight lines on the image plane since the body model is a constructed as a collection of rectangles where each body part is a set of two parallel lines e.g. the two edges of the arm or the torso. The similarity between the image gradient directions and the line orientation along such an edge of some body part is defined as a measure for localizing the body part. An objective function is defined by combining these orientation similarity measures which is maximized through stochastic search iterations [9]. Existing approaches have employed color constancy assumptions, limb detectors along with body part connectivity constrained maximization for pose estimation [2,10,3]. Thus, these approaches are heavily dependent on clothing color constancy, hand skin exposure and are far from real time execution. On the other hand, we have used only the face detector and localized the body parts using image gradient measurements (and hence independent of color constancy assumptions) one by one thereby reducing the dimensionality of the domain of stochastic search leading to real time performance (Section 2). The proposed approach is experimentally validated on a number of human upper body images containing varying levels of background clutter (Section 3). Finally, we conclude in Section 4 and outline the future extensions.

2 Upper Body Pose Estimation

Figure 1 shows the $2D$ human body model (with body part dimensions and joint angles) used in our work. We assume the human body to be near vertical for our particular application domain and hence assume the joint angle between torso and the vertical axis of the head (θ_t) to lie in the interval $[-\frac{\pi}{12}, \frac{\pi}{12}]$. The joint angles made by the left (θ_{lua}) and right (θ_{rua}) upper arms with the torso axis are assumed to vary in the interval of $[0, \pi]$. However, considering the possibilities of roll in the upper arms, the angles between the lower and upper arms at left (θ_{le}) and right (θ_{re}) elbows are assumed to lie in the interval of $[\frac{\pi}{4}, \frac{7\pi}{4}]$. The image co-ordinates of the joints viz. J_{th} (between head and torso), J_{tl} (between torso

and left upper arm), J_{tr} (between torso and right upper arm), J_{le} (left elbow) and J_{re} (right elbow) can be obtained in terms of the body part dimensions and the joint angles using forward kinematics computations [11].

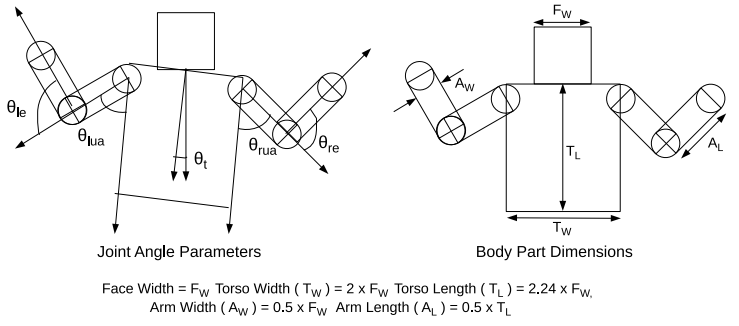

Fig. 1. The human upper body model – The face, torso and arms are modeled as rectangles with circular regions at the joints to generate the effect of smooth contour of the body. The relative length/width of the body parts are derived from anthropometric ratio data [8] in terms of the face width.

Consider the case of localizing a certain body part, e.g. the right upper arm. We first fix the joint co-ordinate (J_{tr}) of the base of the right upper arm. For different values of the joint angles (θ_{rua}), we compute the extent of alignment of the image edges with the outlines of the rectangle representing the right upper arm through an "orientation similarity measure". The final orientation of the right upper arm is obtained at the angle maximizing this measure (Figure 2(a)) and is described next.

Let $m(u, v)$ and $\theta(u, v)$ be the respective gradient magnitude and (unsigned[1]) direction computed from the image pixel position $I(u, v)$. Let the unsigned orientation of a body part outline at the position (x, y) is $\phi(x, y)$. We define the orientation similarity measure as $\gamma(u, v; x, y) = 1 - \frac{|\theta(u,v)-\phi(x,y)|}{\pi}$ ($\theta(u, v)$ and $\phi(x, y)$ being unsigned orientations $\gamma(u, v; x, y) \in [0, 1]$). However, computation of orientation similarity on a single pixel has two major disadvantages – first, the concerned image pixel (u, v) may have a week gradient magnitude indicating lesser importance of the gradient direction; and second, the computation might be susceptible to noise if computed only on a single pixel. Thus, we propose to use a magnitude and position weighted similarity measure ($osm(x, y, r)$) computed over a circular neighborhood $N(x, y, r)$ of radius r around the pixel position (x, y) defined as,

[1] In case of unsigned directions, a line at an angle of $-\alpha$ with the is considered to be equivalent to the line making an angle $\pi - \alpha$ with respect to the same reference axis. The unsigned orientations are thus considered to lie in the interval $[0, \pi)$.

$$osm(x,y,r) = \frac{\sum_{(u,v)\in N(x,y,r)} \gamma(u,v) d(u,v;x,y,r) m(u,v)}{\sum_{(u,v)\in N(x,y,r)} d(u,v;x,y,r) m(u,v)} \quad (1)$$

$$d(u,v;x,y,r) = \begin{cases} 1 - \frac{(u-x)^2+(v-y)^2}{r^2}; & (u-x)^2 + (v-y)^2 \leq r^2 \\ 0; & \text{Otherwise} \end{cases} \quad (2)$$

where $d(u,v;x,y,r)$ is the position weighing function. Let $\mathcal{C}$ be the set of contour pixels of some body part (e.g. the edges of the arms or the torso). We define the net orientation similarity measure $OSM_r(\mathcal{C})$ over the contour $\mathcal{C}$ as $OSM_r(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{(x,y)\in\mathcal{C}} osm(x,y,r)$ where $|\mathcal{C}|$ is the number of pixels in the contour $\mathcal{C}$. We next describe the process of maximizing OSM through stochastic search iterations.

The head region is located first using a Haar feature based frontal face detector [7]. This provides us with F_w from which the body part dimensions are computed using the anthropometric ratio data [8]. This also provides us with the image co-ordinates of the head-torso joint (J_{th}). We perform 3 stochastic search iterations with a population size of 5 angles to estimate the head-torso joint angle θ_t. Localizing the torso provides us with the joint co-ordinates J_{tl} and J_{tr}. To localize the left upper/lower arms, we execute 3 stochastic search iterations with a population size of 10 two-angle $(\theta_{lua}, \theta_{le})$ tuples. A Similar procedure is adopted for localzing the right upper/lower arm. Thus, we need a total of $15 + 30 + 30 = 75$ OSM computations per image at an average of 11.33 frames per second.

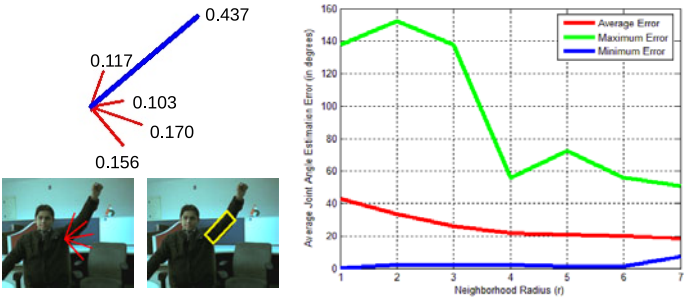

Fig. 2. Orientation Similarity Maximization. (a) The right upper arm modeled as rectangles are placed in different orientations and the direction (shown in blue) maximizes the orientation similarity measure. (b) The minimum/average/maximum joint angle estimation errors for varying sizes of the neighborhood radius. Note that there is not much change in error for values of r exceeding 5.

3 Results

We have performed our experiments on a set of (unrelated) single person (upper body) images downloaded from the web and an image sequence recorded in the

laboratory settings with varying levels of background clutter. A set of 20 images from this data set are ground truthed through manual measurement of the joint angles. We note that the accuracy of pose estimation directly depends on the neighborhood radius r in the computation of OSM. For a certain value of r, we compute the average joint angle estimation error over 5 joint angles from 20 images. Figure 2(b) shows the maximum, average and minimum estimation errors for $1 \leq r \leq 7$. Significant changes in the error is not observed for $r > 5$. However, higher values of r lead to larger number of computations and hence we fix $r = 5$ for our experiments. The results of human upper body pose estimation on 12 images from the data set are shown in figure 3.

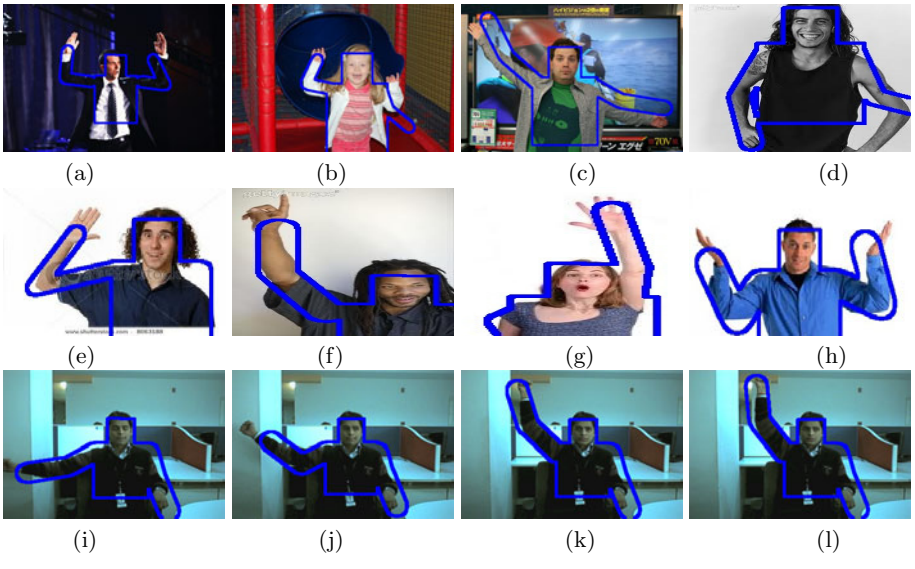

Fig. 3. Experimental results performed on a set of human upper body images with varying levels of background clutter. Although the joint angles are estimated accurately, the human body outline is not fitted tightly on some cases on account of person to person variations in anthropometric data and loose clothing. Failures in localizing in right lower arm is observed in (b) due to loose clothing and strong background clutter (red pipe); where as in (d) the reason is the deviation from the average anthropometric ratio data.

4 Conclusion

We have presented a methodology for human upper body pose estimation from a single image using a $2D$ model (motivated by pictorial structures approach). We have defined an orientation similarity measure to align the body parts (torso and arms) in images. A Haar feature based frontal face detection followed by stochastic search iterations are used to localize the body parts while maximizing

the proposed orientation similarity measure computed along the outlines of the body parts. In contrast to the existing works in human pose estimation from single images, we have achieved high accuracy while performing in real time.

The current work has only focused on the pose estimation from a single image and is a necessary first step towards tracking the pose of the user. We identify the future extensions in two directions. First, the current work is to be extended to tracking where the stochastic search space is further reduced using temporal information. However, it is still not fully possible to estimate the accurate 3D pose from single view information. The second possible extension will be in the direction of combining the image based orientation similarity measurements with depth data obtained from range sensors to avail accurate 3D pose.

References

1. Moeslund, T., Hilton, A., Kruger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding 104, 90–126 (2006)
2. Ramanan, D.: Learning to parse images of articulated bodies. In: Neural Information Processing Systems (NIPS), pp. 1129–1136 (2006)
3. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Pose search: Retrieving people using their pose. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2009)
4. Freifeld, O., Weiss, A., Zuf, S., Black, M.J.: Contour people: A parameterized model of 2d articulated human shape. In: IEEE Conference Computer Vision and Pattern Recognition, pp. 639–646 (2010)
5. Sapp, B., Jordan, C., Taskar, B.: Adaptive pose priors for pictorial structures. In: IEEE Conference Computer Vision and Pattern Recognition, pp. 422–429 (2010)
6. Yang, W., Wang, Y., Mori, G.: Recognizing human actions from still images with latent poses. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2030–2037 (2010)
7. Viola, P., Jones, M.: Robust real-time face detection. International Journal on Computer Vision 57, 137–154 (2004)
8. NASA: Anthropometric Source Book, vol. 2. Springfield VA (1978)
9. Spall, J.C.: Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control. Wiley-Interscience, Hoboken (2003)
10. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
11. Craig, J.J.: Introduction to Robotics Mechanics and Control. Pearson Education Inc., London (2003)