Vision Based Human Pose Estimation for Virtual Cloth Fitting

Sourav Saha
Jadavpur University, Kolkata
souravsaha.juit@gmail.com

Pritha Ganguly
National Institute of Technology, Durgapur
prithaganguly.dgp14@gmail.com

Subhajit Chaudhury
Indian Institute of Technology, Bombay
subhajit.ju4u@gmail.com

ABSTRACT

This paper proposes a real-time solution to setting up a virtual trial-room for on-line portals selling apparels using a generic web camera interface to the portal. The user selects an image of an apparel from the on-line display and captures his/her own videos. The proposed method detects the pose of the user as well as various anthropomorphic features such as length and thickness of upper limbs and the dimensions of the torso. We use a background subtraction based methodology to segment out the human body from the image. The segmented human body contour is represented by a 1D curve by computing the distance of a point on the contour from the body centroid. Various extremities of body parts are found out by measuring the curvature. Using the detected feature points, we use a cloth fitting algorithm to fit the garment to the users body. The entire process is performed at 30fps, providing a realistic rendering of virtual clothing for any user

Categories and Subject Descriptors

I.3.5 [**Computer Graphics**]: Computational Geometry and Object Modeling; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis; I.4.7 [**Image Processing and Computer Vision**]: Feature Measurement

Keywords

Virtual trial-room, generic web camera, cloth fitting, background subtraction.

1. INTRODUCTION

Virtual trial rooms for on-line portals selling apparels have become quite useful in the present scenario with the increasing popularity of on-line shopping among the masses. Buying of apparel at the web portal makes users question their choice which to some extent has led to the abandonment of the shopping cart. These web portals make use of augmented reality which helps them to simulate virtual trial rooms. The trial rooms allow users to visualize themselves in the apparel of their choice before purchasing it and also without physically putting them on. In all the usual cases, the customer stands in front of a screen which shows the real-time visualization of him/her wearing the virtual apparel [6].

Several commercial applications exist in the present market for Virtual Trial Room(VTR) implementation [8]. Styku [2] with the help of its body scanning technology provides size and fit prediction of the chosen garment on the scanned body. The body scanning is implemented using Microsoft Kinect device which creates a complete 3-D model of the user. Similarly, FaceCake's Swivel [3] is labeled as a try-on system that allows users to see how apparels and accessories look on them in real-time. It superimposes a still image of the apparel on the user's body. It also embeds a style advisor which recommends what accessories go well with what garment as the user tries on. It also uses Microsoft Kinect device for its virtual fitting machine. Another augmented reality based 3D fitting room system, Fitnect [1] also uses Microsoft Kinect to provide 3-D garment modelling and rendering to preview products virtually without putting them on physically.

There has been previous works in this field of virtual cloth fitting. An image based approach for virtual cloth fitting described in [10] uses Microsoft Kinect camera to generate a convincing garment simulation. The motion and the deformation of the virtual garment is estimated by tracking the user's skeleton and utilizing the corresponding video stream of the actual garment worn by the user. An image processing method for virtual fitting room described in [8] uses face detection methodology to detect the reference points and uses augmented reality marker to overlay the virtual clothing over the user's image. However, Kinect as a household device is still far from reality and most users do not have any access to Kinect. On the other hand, most household PCs have web cameras that should be used for such an application. However, a lack of 3D data from such cameras makes the problem very difficult.

In our work, we have used a generic web camera which is easily available to capture the video of the user instead of Microsoft Kinect as in [10]. The remainder of the paper is described as follows: Section 2 describes the method of data acquisition from a generic web camera. Section 3 describes the classification of the feature points obtained from the distance curve into head, tip of feet and hand, neck and elbow points. Section 4 describes our cloth fitting algorithm. Section 5 presents our experimental results and Section 6

ICVGIP '14, December 14-18, 2014, Bangalore, India

ACM 978-1-4503-3061-9/14/12 ...$15.00.
http://dx.doi.org/10.1145/2683483.2683543.

concludes the paper.

2. DATA ACQUISITION

The proposed system uses a generic web camera that is usually embedded in all present day laptops for video acquisition. The camera first records the background without the moving foreground object i.e the user. After sometime the user is allowed to be in the field of view of the web camera. Using background subtraction methodology, the human body i.e. the user is segmented out from the video. For t^{th} frame, we can represent our observed image as

$$f_t(x,y) = h_t(x,y) + n_t(x,y) \tag{1}$$

where $f_t(x,y)$ represents the observed image, $h_t(x,y)$ is the original image and $n_t(x,y)$ represents the noise perturbation. A filtering mechanism reduces the noise variance in the background which helps to make our segmentation result precise. The following equation [9] for t^{th} frame is used:

$$h_t(x,y) = \alpha h_{t-1}(x,y) + (1-\alpha) f_t(x,y) \tag{2}$$

where α defined as $0 < \alpha < 1$ represents a constant factor which dictates the importance of the previous and current estimates.

2.1 Frame Difference

The moving object can be easily extracted out by taking the difference between the two consecutive frames, as this is mostly indoor image. For outdoor video, one can use the Gaussian Mixture Model based method as described in [9]. Frame differencing is possible as we have assumed that the background pixels are static and the foreground pixels are moving. The frame difference output is then converted into a binary image by using thresholds [7]. This can be written as

$$FD(x,y)_{res} = |frame(x,y)_n - frame(x,y)_{n-1}| \tag{3}$$

where $frame(x,y)_n$ refers to the n^{th} frame.
In the resultant image frame, pixels with an intensity value greater than or equal to some threshold ($Thres_{fd}$) is given the value 1 and the pixels with values lower than $Thres_{fd}$ are given the value 0. Pixels with the value 1 are more likely to be foreground pixels whereas the pixels with value 0 are more likely to be a part of the background. Thus we have a binary frame difference image.

2.2 Background registration and subtraction

We have assumed that the foreground is not stationary. Hence we need a video input. From the resultant image of the above operation, pixels having the same intensity for a long time are most likely to be the background pixels [7]. A matrix (having the same size as that of the frame) is maintained. The $(i,j)^{th}$ value in the matrix denotes for how many number of frames the intensity of the $(i,j)^{th}$ pixel of the current frame has been stationary. If this value exceeds a certain number, then that pixel is considered to be a part of the background. In this way, the background is recorded and by finding the difference between the current frame and the background, the moving object i.e. the human body can be easily detected and segmented out. Further noise which may still exist in the background mask is removed by refining. At the end of all these steps, we have obtained the background as a black blob whereas the object appears as a white blob.

2.3 Identification of high curvature points

After finding the object mask, 1-D contours of the object is found which is then fitted in a rectangle which best fits its dimensions i.e. all the pixels in the 1-D contour of the moving object lie inside the rectangle. The centre of the mass of the object and the orientation angle of the object is found out. The formula below can be used to be find out the orientation [7] :

$$\theta = \frac{1}{2}\tan^{-1}[\frac{\mu_{1,1}}{\mu_{2,0} - \mu_{0,2}}] \tag{4}$$

where $\mu_{p,q}$ is the central moment with degree (p,q) which can be computed as

$$\mu_{p,q} = \sum\sum (x-\bar{x})^p.(y-\bar{y})^q \tag{5}$$

where (x,y) belongs to the points on the contour and $(\bar{x},\bar{y})$ denotes the centre of the mass of the object.

The midpoints of the sides of the best fitting rectangle are found out and are named as A,D,B and C in the clockwise manner beginning from the top. This would be helpful for pose estimation described later in this section.

Having known the centre of mass and the points of the contour of the foreground object, we calculate the Euclidean distance of each point from the centre and use it to draw a curve known as the distance descriptor. Distance is calculated using the formula:

$$Dist = \sqrt{(x-\bar{x})^2 + (y-\bar{y})^2} \tag{6}$$

. To remove the spikes and erroneous points from the curve, the distance descriptor is filtered using the formula:

$$DLow_i = \frac{1}{2*n+1}\sum_{j=-n}^{n} Dist_{i+j} \tag{7}$$

where 'n' can be any positive integer which signifies the span of the number of regions which we want to average in order to smooth the distance descriptor.
To locate the head, tips of hand and feet, we need to locate the high curvature points i.e. the local maxima on the filtered curve. Local maxima is found by calculating the gradient of the points on the curve. Gradient of any point on a curve can be found by differentiating the curve with respect to x co-ordinate at that point.
Local maxima points will be those points which will have points with negative gradient to the left of it and points with positive gradient to the right of it. The curvature of each of the high curvature points denoted by (x_c, y_c) is calculated using the formula [7] :

$$\theta_i = \cos^{-1}[\frac{u.v}{mag(u).mag(v)}] \tag{8}$$

where $u.v$ denotes the dot product between the vectors u and v for the i^{th} high curvature point. The vectors u and v are estimated as:

$$u(x_f, y_f) = (x_{c+k}, y_{c+k}) - (x_c, y_c) \tag{9}$$

$$v(x_f, y_f) = (x_{c-k}, y_{c-k}) - (x_c, y_c) \tag{10}$$

Here the point (x_{c+k}, y_{c+k}) represents the point on the 1-D contour whose x co-ordinate and y co-ordinate is 'k' more than the respective co-ordinates of the high curvature point. Similarly, (x_{c-k}, y_{c-k}) represents the point on the 1-D contour whose x co-ordinate and y co-ordinate is 'k' less than

the respective co-ordinates of the high curvature point. The value of 'k' can be any positive integer(we have considered 'k' to be 20). The θ_i obtained is always greater than 0 and is less than or equal to π.

3. POSE ESTIMATION

For pose estimation of the segmented human body, the midpoints of the rectangle A,D,C and B (as discussed in the previous section), the orientation of the body and curvature of the high curvature points are estimated [5].

3.1 Identification of Head, tip of Feet and Hands

To locate the head, first the search region is estimated. As the bounding rectangle best fits the human body, it has been observed that the head is likely to be located in the upper half of the rectangle i.e above the line joining C,D and below the point passing through A, when the orientation of the body i.e $\theta >= 15^\circ$ when the user is in the standing posture and his head is always above his feet. In this search region, the high curvature points are found out. The following cases may occur:

1. If only one high curvature point is obtained, then that point is classified as the head.
2. If two points are found, then their curvature angles θ_i are compared. The point which satisfies the angle constraint is selected as the head. If none or both of the points fulfil the criterion, then the one with the smaller curvature angle is selected and marked as the head.
3. If no point is found in the search region, then point A is selected and marked as the Head.

To locate the tip of feet, the search region estimated is the lower half of the rectangle i.e below the line joining C,D and above the line passing through B. High curvature points are found in this region. The following cases may arise:

1. If two points are found and they satisfy the angle constraint, then those two points are classified as feet.
2. If more than two points of interest are found, then the two points which have the maximum distance from the head point, is considered as the tip of feet and is marked.

The remaining unmarked high curvature points are likely to be the potential points for the tip of hands. If more than two points are found unmarked, then among them, the two points which have maximum distance from the centre of mass is classified as hands. If no points are found, then automatically points C and D are marked as tip of hands.

3.2 Identification of Neck and the Elbows

To begin with, we first try to locate the neck point on the 1-D contour. It is geometrically observed that the neck point divides the line joining the tip of the head and the C.G in the ratio 1:2 as in [5]. With the help of the neck point, the left and right elbow points can be estimated.

The left elbow points denoted by $\{E_{L1}, E_{L2}\}$ and the right elbow points denoted by $\{E_{R1}, E_{R2}\}$ is approximately near the midpoint M_L and M_R of the line joining the neck point and the tip of left and right hand respectively. Depending on where M_L and M_R lies, the following cases may arise:

1. For the left hand, if the midpoint M_L lies within the contour, a line inclined at a certain estimated angle (we have taken the angle to be approximately 90°)is drawn through the point which intersects the contour at two points. The intersecting point on the contour with the minimum curvature, is considered to be one of the elbow points (E_{L1}). Depending on where the above elbow point is obtained i.e in the upper or lower contour, a perpendicular line can be drawn from that point to the lower or upper contour in order to locate the other elbow point (E_{L2}).
2. For the left hand, if M_L lies outside the contour, then again a line inclined at a certain estimated angle is drawn which intersects the contour at two points. The one with the minimum curvature is chosen as one of the elbow points (E_{L1}). In the same way as described above, the other elbow point(E_{L2}) is found out.

4. CLOTH FITTING

The feature points on the 1-D contour thus obtained can be used to draw a time varying stick figure,using a time-varying directed acyclic graph (DAG),$\mathbf{G(t)=(E(t),V(t))}$, with its directed edges $\{e_k(t) \in \mathbf{E(t)}\}$, representing the bones in the body and the vertices $\{v_i(t) \in \mathbf{V(t)}\}$ representing the joints. The time varying nature of $\mathbf{G(t)}$ is not due to change in edge connectivity with time but due to changing vertex locations and possible occlusion of some joints.

For the purpose of cloth fitting [4], we first try to superimpose the cloth on a human template model. We ask users to stand in front of the web camera in a specified posture (we used the T-pose as the initial posture for our purpose) so that the initial cloth fitting can be done in this posture itself. For this initial fitting of the upper garment, we fit three rectangles $Rect_{lhand}, Rect_{rhand}, Rect_{body}$ for the left hand, right hand and body respectively. As the body moves, the orientation of the edges vary with time and accordingly we fit the rectangles so that the cloth fits precisely on the body.

For the hands, the width of the rectangles $Rect_{lhand}$ and $Rect_{rhand}$ can be obtained by calculating the distance between the two elbow points for either hands($\{E_{L1},E_{L2}\}$ for left hand and $\{E_{R1},E_{R2}\}$ for right hand). The length of the above rectangle is the length of the edge which joins the neck point and the midpoint of the elbow points for either hand.

For the body, the length of the rectangle $Rect_{body}$ is equal to the length of the edge between the neck point and the centre of mass while the width of the rectangle is obtained by calculating the distance between two points on the contour which are again obtained by drawing a horizontal line through the centre of mass(points obtained are the torso points). The user may want the cloth to be either skin fitting or loose fitting. For the former, the length of the best fit rectangle $Rect_{body}$ is calculated as discussed above. For the latter, the length of $Rect_{body}$ is increased by an amount α, where α is a positive constant which depends on the size of the user.

To find the orientation $\theta_k(t)$ of the edge at any time 't' representing the hands, we use the formula:

$$\theta_k(t) = tan^{-1} \frac{y_{elbow}(t) - y_{neck}(t)}{x_{neck}(t) - x_{elbow}(t)} \tag{11}$$

where (x_{elbow}, y_{elbow})represents the midpoint of the two elbow points of each hand and (x_{neck}, y_{neck}) represents the neck point.

After initial cloth fitting is complete, we need to mould the edges of the garment around the body to give an impression to the user that he/she is actually wearing that garment. To fit the garment perfectly, some post processing is required.

The points of the rectangles($Rect_{lhand}$,$Rect_{rhand}$,$Rect_{body}$) which lie inside and/or on the 1-D contour are considered to

be the points on the garment. The points of the rectangle lying outside the contour are considered to be extra pieces of the cloth and are discarded.

5. EXPERIMENTAL RESULTS

The proposed method has been implemented on a computer which runs Windows 7 operating system with i7-2600 CPU @ 3.40Ghz, 16GB RAM and NVIDIA GeForce GTX 560i GPU. Microsoft Visual C++ 2010 with OpenCV library is used for the required segmentation and cloth fitting.The frame width and height of the video is 640 and 480 respectively and the frame rate is 30 frames/second assuming that the web camera is stationary.

In Section 2.2, the number of frames for which the pixels were observed before classifying them as background is taken to be 80 frames in our experiment. In Section 2.3, while filtering the distance descriptor, the span of the number of regions which we want to average in order to smooth the curve is taken to be '13' i.e for our purpose, we have taken 'n' to be '13'. In Section 3.1, the curvature angle constraint for identifying the head is observed to be less than 179°. For identifying the tip of feet, the curvature angle of the high curvature point has to be less than 100°. For identifying the tip of hands, there's no specific angle constraint that needs to be satisfied.

A snapshot of the initial video taken by the user is shown in Figure 2. This figure describes that the initial T-pose which we have used for calibration. The object mask of the user that is segmented out from the background is shown in Figure 3. The outline of the object mask obtained is not smooth due to the noise present in the background of the initial video. In Figure 4, the white blob represents the foreground object which is segmented from the initial video. In Figure 6, the 1-D contour of the foreground object i.e the user is shown, where the point N indicates the estimated neck of the user,points E1 and E2 corresponds to the upper and lower elbow points and two extremities of the hand points are marked as Limbs. After obtaining and classifying the feature points, the chosen garment is fitted on the contour using the cloth fitting algorithm in Section 4. Figure 5 demonstrates the initial fitting of the virtual cloth (Figure 1) on the user. After post processing, the precisely fitted apparel on the user's figure can be seen in Figure 7 and 8.

Our system has certain shortcomings. In the poses where the hands of the user is below the torso, it becomes difficult to estimate the limb points. Hence, the cloth cannot be fitted properly as shown in Figure 9. Similar difficulty is observed with Microsoft's Kinect device too as when the hands of the user is below the torso, the depth of the limb points coincide with the depth of the torso points making it difficult to be estimated precisely.

6. CONCLUSIONS

In the proposed system we attempt to provide a real-time solution of setting up a VTR for the on-line portals selling apparels using a generic web camera. The same can be done with the help of Microsoft Kinect device which, although cheap is not available to all the general users. Thus this algorithm uses web camera which are embedded with the present day laptops to provide a cheaper interface to the virtual world of trial rooms. This also makes our proposed system commercially viable.

Figure 1: A sample apparel which is to be fitted.

Figure 2: A snapshot of the video taken by the user.

Figure 3: The object mask.

Figure 4: The segmented human body from the video.

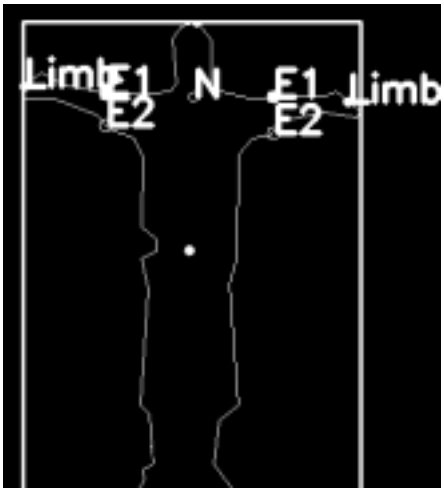

Figure 6: The 1-D contour is fitted in a bounding rectangle and the feature points are marked on the contour where 'E1 and E2' represents the Elbow points, 'N' represents the neck point.

Figure 5: The user with the virtually fitted apparel

Figure 7: The user with the virtually fitted apparel after post processing

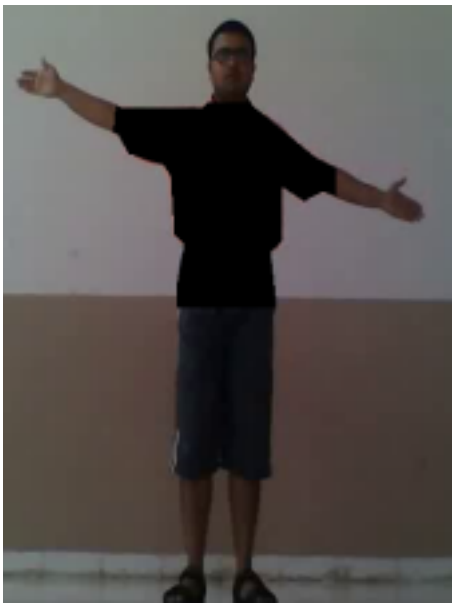

Figure 8: The user with the virtually fitted apparel in a different pose

Figure 9: Demonstrating the shortcomings of our system

In our proposed work, we have assumed the web camera to be stationary. Virtual cloth fitting with moving camera is the task which is yet to be explored. In our proposed system the user only moves horizontally. Vertical movement towards or away from the camera, or sideways posture is not permitted as image depth analysis using the web camera is not incorporated in our proposed system. Incorporation of the same would provide a good competition to Microsoft Kinect device as the same result can then be obtained using a camera which is easily and cheaply available nowadays. The texture rendering of the apparel onto the segmented body is also not attempted here. In our future work, we will eventually try to implement the same. We are also trying to increase the accuracy of the implementation, but the main strength of the method lies in the cloth fitting using 2D images without any depth cues for cloth fitting.

7. ACKNOWLEDGEMENTS

The authors wishes to express their sincere gratitude to Professor Subhasis Chaudhuri, Department of Electrical Engineering, Indian Institute of Technology, Bombay for his valuable advice and guidance throughout the course of this work.

8. REFERENCES

[1] Fitnect, http://wiki.fitnect.hu/.

[2] Styku, 2011: http://www.styku.com/business.

[3] Swivel, 2011. http://www.facecake.com/swivel/index2.html.

[4] M. Aono, P. Denti, D. E. Breen, and M. J. Wozny. Fitting a woven cloth model to a curved surface: Dart insertion. *IEEE Comput. Graph. Appl.*, 16(5):60–70, Sept. 1996.

[5] T.-C. C. Chia-Feng Juang and W.-C. Du. Human body 3d posture estimation using significant points and two cameras. *The Scientific World Journal*, 2014.

[6] S. Gray. In virtual fashion. *IEEE Spectr.*, 35(2):18–25, Feb. 1998.

[7] C.-F. Juang, C.-M. Chang, J.-R. Wu, and D. Lee. Computer vision-based human body segmentation and posture estimation. *Trans. Sys. Man Cyber. Part A*, 39(1):119–133, Jan. 2009.

[8] C. G. Martin and E. Oruklu. Human friendly interface design for virtual fitting room applications on android based mobile devices. *Journal of Signal and Information Processing*, 3(4):481–490, 2012.

[9] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2, pages –252 Vol. 2, 1999.

[10] Z. Zhou, B. Shu, S. Zhuo, X. Deng, P. Tan, and S. Lin. Image-based clothes animation for virtual fitting. In *SIGGRAPH Asia 2012 Technical Briefs*, SA '12, pages 33:1–33:4, New York, NY, USA, 2012. ACM.