Computational Visual Media
DOI 10.1007/s41095-017-0084-6

Research Article

Image-based clothes changing system

Zhao-Heng Zheng[1], Hao-Tian Zhang[2], Fang-Lue Zhang[3] (✉), and Tai-Jiang Mu[4]

Abstract Current image-editing tools do not match up to the demands of personalized image manipulation, one application of which is changing clothes in user-captured images. Previous work can change single color clothes using parametric human warping methods. In this paper, we propose an image-based clothes changing system, exploiting body factor extraction and content-aware image warping. Image segmentation and mask generation are first applied to the user input. Afterwards, we determine joint positions via a neural network. Then, body shape matching is performed and the shape of the model is warped to the user's shape. Finally, head swapping is performed to produce realistic virtual results. We also provide a supervision and labeling tool for refinement and further assistance when creating a dataset.

Keywords clothing try-on; image warping; human segmentation

1 Introduction

With the rapid development of the Internet and digital media, current graphics tools fail to meet the increasing demands for personalized visual content manipulation. Changing clothes in images captured by users is one such application. However, using current image editing technology in professional image editing packages such as Adobe Photoshop, changing clothing in images is still a labor-intensive and time-consuming task. Realistically changing clothing has many challenges, such as the deformation of the clothes in 2D images, differences in luminance conditions, clothes region composition, and so on. It is hard even for experienced artists to generate good results. Moreover, in order to produce a visually realistic image, artists must fit target clothes to target bodies exactly, which requires considerable effort, using perhaps hundreds of local edits.

Researchers have made efforts to find easier ways for users to perform this task. Instead of fitting clothes to target bodies, Chen et al. [1] proposed a clothes swapping algorithm to address changing clothes having a single color with others having similar styles, based on human body retrieval results. However, it is not applicable to a freer and more complicated task: changing clothes for ones with different textures and styles. Zhou et al. [2] proposed a parametric body reshaping system with manually adjusted parameters, offering a way of fitting body size to clothes. However, their method did not work well on natural images, as more consideration of the final composition results is needed for the natural images provided by users.

In this paper, we propose an automatic clothing changing system. We allow a user to input his or her own image, and choose clothes from models in our image library. We first segment the input image and library image to obtain a precise mask of the human body. Then joint detection and boundary matching is performed. After matching the two bodies, we calculate their differences in a parametric way, as a basis for the following content-aware image warping

1 Computer Science and Engineering, University of Michigan, 2260 Hayward St, Ann Arbor, MI 48109, USA. E-mail: zhaohenz@umich.edu.

2 Computer Science Department, Stanford University, 353 Serra Mall, Stanford, CA 94305, USA. E-mail: zhanghaotiansola@gmail.com.

3 School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zealand. E-mail: z.fanglue@gmail.com.

4 TNList, Tsinghua University, Beijing 100084, China. E-mail: mmmutj@gmail.com.

Manuscript received: 2017-03-17; accepted: 2017-04-09

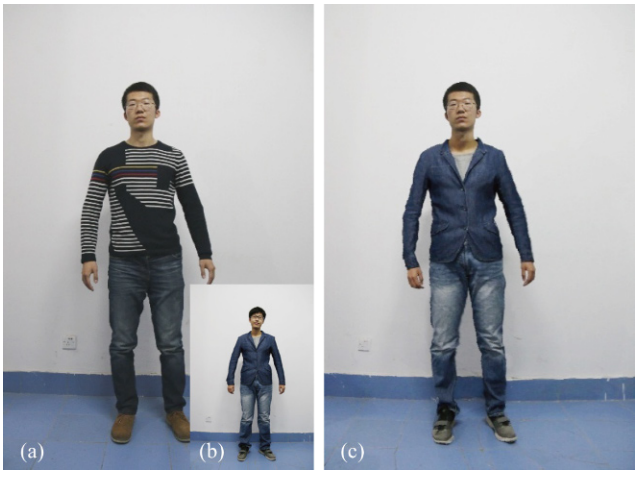


Fig. 1 Example input and output: (a) the user input, (b) model image selection, and (c) the result.

stage, which warps the target body into the desired shape. Head swapping [1], the last step, is then applied. As shown in Fig. 1, a synthesized photo of the user with correct body shape and desired clothing is produced.

2 Related work

2.1 Articulated human pose estimation

The development of articulated human pose estimation greatly boosted semantic understanding of images. Ferrari et al. [3] proposed a reduced search space algorithm, which improved the chance that estimation would succeed. Andriluka et al. [4] introduced a generic approach based on pictorial structure, while Johnson and Everingham [5] combined color segmentation and limb detection. Part-based models [5–8] have proved their effectiveness. Vineet et al. [9] introduced detector-based conditional random fields. Combination of convolutional neural networks (CNN) and graphical models [10] has also achieved high performance. Taking a different approach to part-based graphical models, Ramakrishna et al. [11] proposed an inference machine with rich spatial interactions. Carreira et al. [12] put forward iterative error feedback, improving the performance of hierarchical feature extractors. Recently, Wei et al. [13] combined CNNs and a pose machine [11], giving a system with state-of-the-art performance.

2.2 Semantic image segmentation

Semantic image segmentation has been one of the most popular topics in recent years. Blake et al. [14] proposed an interactive method with probabilistic formulation. With the goal of creating an automatic system, conditional random fields (CRF) were applied. Further work [15, 16] has focused on CRF, producing desirable results. In the meantime, other work [17–20] has adopted region grouping to address this problem. Furthermore, hierarchical segmentation models [16, 18, 19] have been combined with the abovementioned methods to improve segmentation performance. Salient region detection [21, 22] and attention models [23] have also greatly contributed to enhancement of image segmentation. Zheng et al. [24] proposed a machine learning method with pixel-wise labels of object class and visual attributes. Recently, with the development of deep learning, neural networks have demonstrated their strong power for semantic image understanding. Recent neural network models [15, 23] provide state-of-the-art image segmentation systems.

2.3 Image warping

There are many image warping algorithms to manipulate an image using control handles. The deformation of the image to follow those control handles can be achieved by optimization methods [25], radial basis functions (RBF) [26], or moving least squares (MLS) [27]. Recently, Zhou et al. [2] proposed a resizing method warping a human body to follow a body skeleton, controlled by varying body parameters. Kaufmann et al. [28] put forward a single unifying framework for image warping based on a finite element method (FEM).

3 Overview

A clothes changing system requires body shape perception and content-aware image warping. Existing methods ask the user to manually input their body parameters to provide information about body shape. In contrast, we present an automatic matching and reshaping system that utilizes a deep neural network to detect articulation, and produces a set of key points to guide shape matching and image warping. See Fig. 2 for an overview of our system.

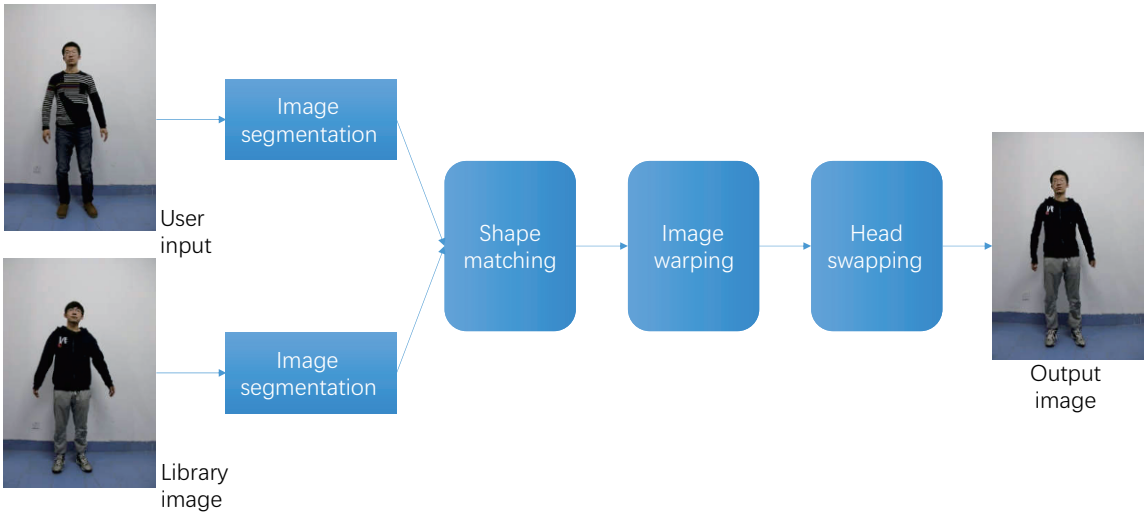


Fig. 2 Overview of our system. Images are first segmented to produce body shape masks. Shape matching is then performed between the two shapes, which guides the following image warping and head swapping processes.

We adopt GrabCut [29] as our basic image segmentation algorithm. It requires prior knowledge to guide the segmentation process. We provide an initial heatmap generated by saliency filtering [30] as well as skin detection [1] to the algorithm. After segmentation, a guided image filter [31] is applied to the mask, refining the human contour. In most cases, it works well and produces suitable segmentation results. In order to deal with difficult cases, we provide a user interface which allows the user to preview the mask, and interactively refine it if needed.

Next, joint detection is performed using a convolutional pose machine (CPM) [13]. We first use a boundary scanning algorithm to extract a set of control points along the body contour. Next, we match the control points of user and library images, according to joint correspondence previously obtained. Then we generate meshes on the target image using constrained Delaunay triangulation. Mesh-based image warping is performed to synthesize the new photo. Finally, head swapping is performed using the method in PoseShop [1] with Poisson fusion [32], to transfer the user's head to the reshaped model's body. A synthetic image of the user wearing the model's clothes is the final output of our system.

4 Automatic mask generation

In this section, we present technical details of our image segmentation method and begin with GrabCut [29], an image segmentation tool.

GrabCut is a powerful algorithm based on graph-cut. It performs well with user interaction and iterative mask optimization. However, as our goal is an automatic system, we have to provide the system with prior knowledge produced by other methods. In order to better guide GrabCut, we determine high-saliency regions [30], in the form of a heatmap H, and turn the heatmap into a four-valued indicator function using a set of thresholds:

$$I_{\text{salicency}}(i,j) = \begin{cases} 0, & H(i,j) \in [0,70) \\ 1, & H(i,j) \in [70,120) \\ 2, & H(i,j) \in [120,180) \\ 3, & H(i,j) \in [180,255] \end{cases}$$

The values 0 to 3 respectively represent background, probable background, probable foreground, and foreground. Generally, this information is sufficient for GrabCut to produce a good segmentation. However, in some cases, it needs further semantic information to do so as the saliency model sometimes does not correctly predict the saliency of body parts. In order to deal with this problem, we tried to augment the differences between regions within an image, but this did not provide sufficient extra information either. Since we are specifically segmenting a human, we should take the most distinctive parts of the human body into consideration. Poseshop [1] uses a skin detector based on a Gaussian mixture model, which produces accurate skin heatmaps of images. Our experiments on segmentation indicate that combining these two

heatmaps works well. We denote the binarized skin heatmap as I_{skin}. We combine the two heatmaps into a fused indicator I, which is

$$I(i,j) = I_{\text{salicency}}(i,j) \vee (3\, I_{\text{skin}}(i,j))$$

This indicator allows us to separate the human from the background, and generate a preliminary mask of the desired regions. Due to the lack of consideration of color, some textures inside the clothing may be still recognized as background. To address this problem, we apply hierarchical contour detection, obtaining a polygon indicating the human contour. Regions inside the polygon are considered to be part of the human body.

Although GrabCut successfully segments the image based on our specific semantic information, the edge of the mask is still rough, containing many noise pixels that degrade matching accuracy later. To refine the mask, we considered applying median filtering and weighted image matting [33]. While weighted image matting removes noise pixels, it is quite time-consuming, taking more than 1 minute to process an image of resolution 480×640. Thus, instead, we use a guided image filter [31], which significantly accelerates the matting progress, producing a refined mask in 0.1 s on an Intel i5 PC with 4 GB RAM using a single thread. Figure 3 demonstrates how our method works.

5 Partial shape matching

Matching the contours of two bodies is the next task to address. Given the mask obtained from the previous section, we extract a set of boundary points (denoted by C) from the body contour using an edge detection algorithm. We wish to *accurately* match two such sets of contour points from different persons: small errors in matching could introduce obvious distortions when performing image warping.

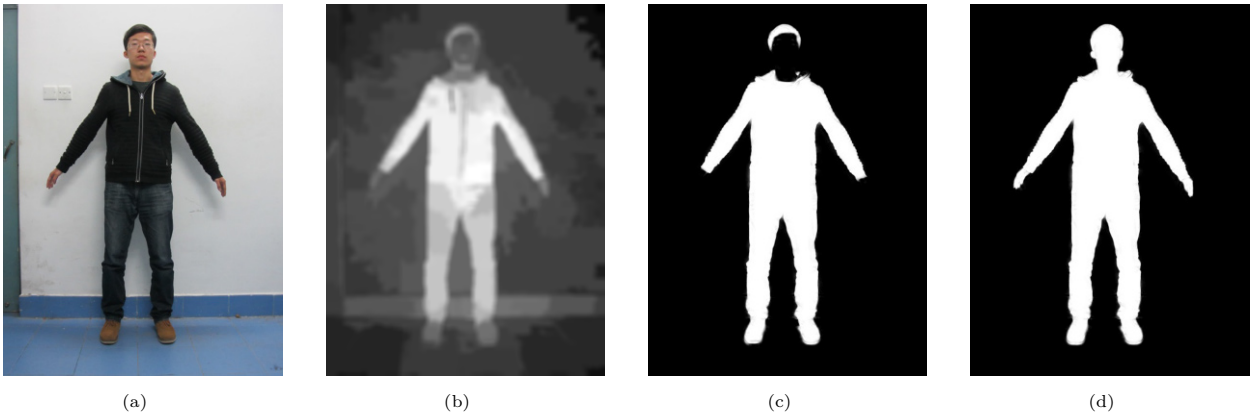
A straightforward method for shape matching is to use shape contexts [34]. However, it has two drawbacks in our task.

Firstly, the human body contour lacks distinctive shape information, usually resulting in cross matching (see Fig. 4(a)) and part shifting (see Fig. 4(b)). Even attempting to match body part contours using this approach still fails to provide satisfactory matching results.

Another handicap of utilizing shape contexts [34] is that it is not a body-specific method. Without correct use of semantic information, we will have difficulty in image warping and head swapping processes. An inelegant way of overcoming the problem is for the user to label key points along the body contour. Other points' locations can be estimated according to their positions relative to these key points. However, these points must be labeled for each input image and library image, which is a labor intensive process.

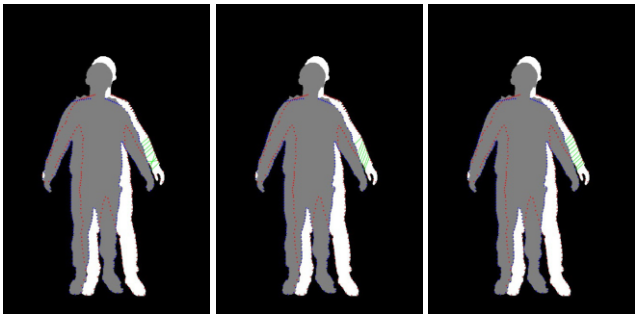
Instead, inspired by articulated pose estimation, we adopted another solution to this problem. Recent works on pose estimation provide human skeletons or key joint points as their output; the latter can provide vital key points for our matching task. We adopt convolutional pose machines (CPM) [13], a state-of-the-art pose estimation system, to provide joint detection for our system. For each input image, we obtain 14 joint points J, as shown in Fig. 5(a). J contains joints at head, neck, shoulder, elbow, wrist, hip, knee, and ankle.

Utilizing joint information, we perform body-aware control point selection from the contour points C. Let the set of control points be P. For each $p \in P$,



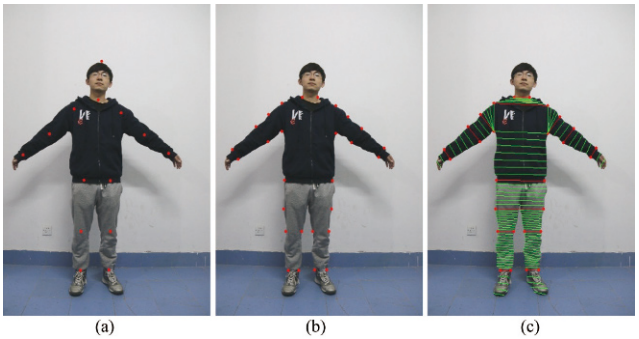
(a) (b) (c) (d)

Fig. 3 Mask generation. We first extract a saliency heatmap (b) from the input (a); this heatmap lacks some body parts (c). A skin detector is used to overcome this problem, and the final mask (d) is optimized by median filtering and guided image filtering.



(a) (b) (c)

Fig. 4 Failure of matching using shape contexts [34]: (a) cross matching, (b) part shifting, and (c) correct matching.



(a) (b) (c)

Fig. 5 Joints and control points: (a) joints obtained using CPM, (b) control point pairs P, and (c) control point pairs Q.

p denotes a pair of control points located on the body contour: see Fig. 5(b). P also contains 14 pairs of points, each pair being calculated from one or two nearby joints in J. Most pairs are directly obtained by calculating intersections of the line perpendicular to the skeleton with C, such as the pairs on the arms and legs. For some joints, such as the shoulder and armpit, we also make use of gradients to help locate a precise control position.

The final step is to obtain denser control points (denoted by Q) along the contour for triangle mesh generation. In order to maintain body semantics, we interpolate between neighboring control pairs in P: see Fig. 5(c).

In the whole process, control points are numbered in a fixed order relative to the body. Thus, the final control point pairs in Q are strictly formatted, so that two such sets from different people are in correspondence. This method is not influenced by differences in body pose, as each control point pair is calculated only using body part information.

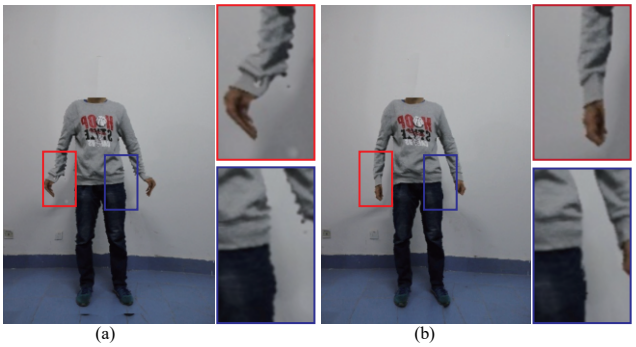
6 Mesh-based image warping

In this section, we introduce our image warping method which transfers the body shape from the user to the model. The control points obtained from the user and the model will be denoted by Q^* and Q respectively.

We tried two standard image warping methods, using radial basis functions (RBF) [26] and moving least squares (MLS) [27]. They both take as input two sets of control points in one-to-one correspondence, so we can directly use Q^* and Q as input. Then, they map the positions of Q to the positions of Q^* to induce an image warp, allowing propagation of the shape differences between the sets of control points to the whole images.

Two warping results are shown in Fig. 6. Our experiments show that MLS [27] tends to generate distortion and aliasing along the body contour. In Fig. 6(a), distortion on both hands and small oscillations on the inside of both arms can be clearly seen. It appears that this is because control points here are too close. Specifically, if a pixel p along the inside of the left arm is close to a control point c, its warped position p^* will be dominated by c, so it will stay close to c^*, the warped position of c. However, the warped position q^* of a pixel q between two control points c_1 and c_2 will be affected by control points on the side of the body, thus causing it to float in or out.

Instead, RBF [26] warping tends to perform better locally, and hardly generates any visible distortion in the warped image: see Fig. 6(b). Thus, we adopt the RBF warping method. Warping is performed as follows. Firstly, we divide the image into a 2D triangular mesh M by applying constrained Delaunay triangulation to the control points Q on the body contour and several points selected on the edge of the image. Then, for each triangle $t \in M$, we get its current vertices from Q and also the corresponding warped vertices from Q^* according to



(a) (b)

Fig. 6 Body shape warping results: (a) using MLS, (b) using RBF.

the matching, giving the warped triangle t^*. Finally, we calculate the position of each pixel in t from its barycentric coordinates in t^*, and obtain its color value by bilinear interpolation.

Given the precise matching result from the previous section, this allows our warping method to accurately transfer the shape information from the user to the model. Even in extreme cases with obvious differences in height or weight, realistic warped results can be synthesized using our method.

The whole warping process takes less than 1 s using an Intel i5 PC with 4 GB RAM, using a single thread.

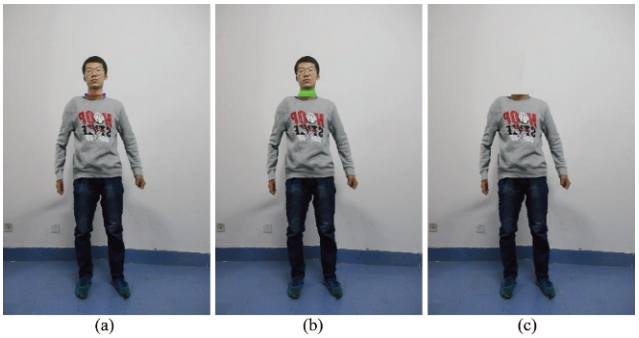
7 Head swapping

Previous sections concentrated on reshaping. However, a similar body shape is still far from our final objective, as people are recognized mainly by their faces. Therefore, we need to perform a head swap as the last step towards our goal.

Chen et al. [1] proposed a method for head swapping. Since we only perform head swapping between two frontal face, we do not need face inputs from other directions. We can directly extract the head contour from the previous mask as the target and source inputs. Our method has the following four steps.

First we extract the head contour from the previously obtained mask; we also use the skin detection [1] result in case the person wears a hood. In this case, the mask will include the clothes near the neck, and only by applying skin detection can we obtain correct neck contour.

We then determine the stitch line and fusion area. The stitch line (the red line in Fig. 7(a)) comes from the control point pair on the neck (blue points in Fig. 7(a)). The area containing the stitch line on the neck (the green area in Fig. 7(b)) is the fusion area; its height is 1/6 of the length of the stitch line. Given the precise control point pair on the neck from matching and exact image warping along the neck, it is straightforward to locate the stitching line, so we can avoid the complicated search for an appropriate stitching line required in Poseshop [1]. Besides, as the neck has nearly the same width after image warping, the stitched result matches the body well.



(a) (b) (c)

Fig. 7 Head swapping. (a) Blue: control points on neck, red: stitching line. (b) Green: Poisson fusion area. (c) Result after removing the original head from the reshaped body.

We next replace the head. Let the area above the stitch line and limited by the head contour be H. We fill the area H with the values of the pixels which are horizontally symmetric with respect to the head contour: see Fig. 7(c). Then we directly stitch the head area H^* from I^* into the warped image, with matting along the boundary.

Finally, we perform Poisson fusion [32] on the fusion area to hide the stitching line.

8 Results and discussion

8.1 Discussion

Further matching results are shown in Fig. 8. Our method can generate uniformly good matching results for different kinds of input pose. Thanks to CPM [13], we are able to detect key joints which define the boundaries of limbs so that we can match control points appropriately.

Some final results are shown in Fig. 9. The input images have resolution 480×640. The first row shows photos captured by the user, while the second row shows model photos sampled from an image library we collected. Results are shown in the bottom row. In these results, our method is good at maintaining body shape from the user input. Specifically, our system can precisely evaluate the shape differences between two people, especially if they are wearing tight-fitting clothes, like the user in Fig. 9(b). Careful choice of control points in our method ensures that deformations of human body parts can be limited to their own regions. See Fig. 9(c) for example, where although the model is fatter, our approach successfully reshapes body parts separately. Additionally, our system also performs body-wise reshaping, so, e.g., in Fig. 9(d), we successfully reshape the model to a different height. Figure 9(g) shows an extreme case in which the user

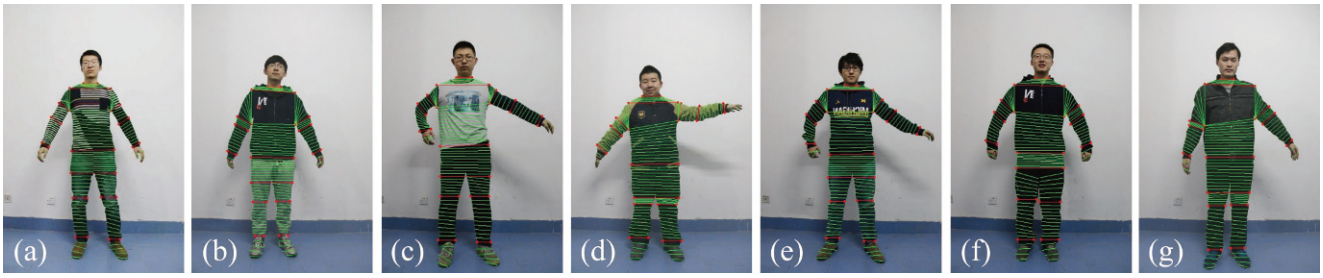


Fig. 8 Matching results for our system.

Fig. 9 Results for our system.

differs a great deal in both weight and height from the model, but our result still accurately preserves the shape of the user.

As well as shape difference, our method can handle differences in pose between user and model, since we have precise matches. Figure 9(f) illustrates that our system works well.

8.2 Comparison

Shilkrot et al. [35] put forward a system for clothing changing and identity transfer. Their method also employs a graph-cut-based method to segment the human body from the background; they utilize user interaction to correct the result iteratively. They differ in using parametric polycurve fitting as the initial guess for the graph-cut segmentation, while we adopt skin and saliency detection. Comparative results show that our initialization provides good prior knowledge allowing generation of a better segmentation without interaction, simplifying the procedure.

Both methods use Poisson blending for head swapping, or so called identity transfer, and it proves its power. With the help of a convolutional pose machine (CPM) [13], we can precisely calculate the position of the neck and automatically resize the head, rather than requiring user guidance, as in Shilkrot's method. Thus, our method needs less user

interaction than their method.

In body warping, our method is quite different from their method, which only asks for a head photo as input, while we use a picture of the whole body. In order to estimate the warping parameters, Shilkrot et al. [2] used a parametric body model to fit the input, which is robust but less accurate. Our method instead takes the whole body as input and warps the target image in an accurate way, reflecting the body differences between user and model. Taking Fig. 10 as an example, we show that our method can appropriately reshape the model to different height and weight according to the input image, while their result is less accurate with respect to body shape.

8.3 User study

In order to evaluate our system, we invited 8 users to give their opinions. We selected 8 models and asked them to score the synthesized photos from 1 to 5 accordingly (1 for very bad, 2 for bad, 3 for so so, 4 for good, and 5 for perfect). We achieved an average score of 3.78, as detailed in Table 1. In most cases, the average score from each user exceeded 3.5, demonstrating that most users were reasonably satisfied with our clothes changing results.

8.4 Limitations and future work

Our system has some limitations. Firstly, success of color-space-based image segmentation depends heavily on brightness and illumination. In Fig. 11(a), the color of the T-shirt is so similar to that of the wall

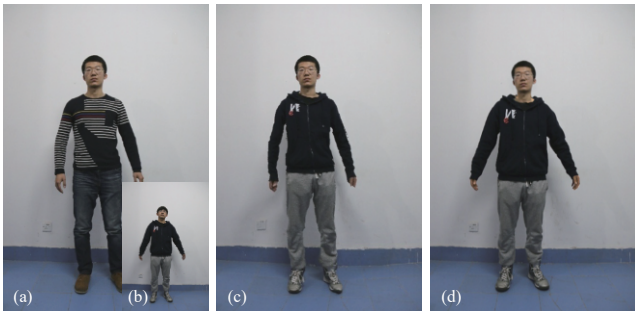


Fig. 10 Comparison between our method and that of Shilkrot et al.: (a) user image, (b) library image, (c) our result, and (d) their result.

Table 1 Average scores awarded by users

User	Average score	User	Average score
1	3.50	5	3.13
2	4.25	6	3.75
3	4.13	7	4.13
4	3.38	8	4.00

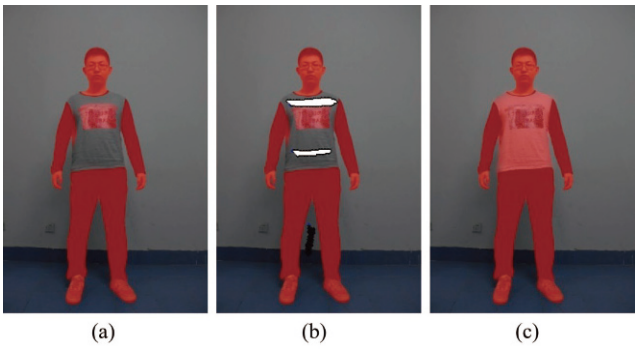


Fig. 11 User interface: (a) initial segmentation failure, (b) user adjustment, and (c) refined result.

that our detector cannot segment it from the wall. To overcome such problems, we provide a window where the user can manually indicate pixels' labels. In Fig. 11(b), we demonstrate how this works. The white region is labeled as foreground, while the black region is labeled as background. With the help of extra information, refinement is achieved, providing a suitable mask as in Fig. 11(c).

We hope in future to collect more data and create a neural network which can segment an image into separate body parts such as arms and legs.

Secondly, our image segmentation algorithm only works when there is no overlap between limbs and body. We may solve such cases by requiring the user to denote the overlapped area. A simpler approach is to require there to be no overlap in the input, which is not especially restrictive.

Thirdly, our system only relies on the contour of the clothes, not the real human body shape itself, for shape recognition. Therefore, the system tends to output better results when users wear tight-fitting clothes. A person wearing thick clothes will be considered to be fat. Unfortunately, we cannot determine correct human body information from a single image.

Finally, while shape matching is usually correct, the current warping method is rather naive, occasionally causing distortion in hands and shoes. Our results show that RBF warping tends to generate distortion where control points are sparse, such as hands and ground lines. Generating a dense mesh for the whole image may contribute to this problem, but matching these extra vertices for warping is another problem, which we leave as future work.

9 Conclusions

In this paper, we have presented a system to help users change their clothes based on their own photo and an image library. Given a photo of a user as input and a model image from the library, our system can automatically synthesize an image showing the user wearing the model's clothes. Our system has broad applications. For example, it could be developed into an online application, where anyone can upload their own photos and choose clothes they want to try on. Furthermore, our system could be utilized in a clothes store setting combined with a camera system, where customers can take a photo of themselves and choose the clothes they like in the store to see what they would look like wearing them. The total process would take no more than 1 minute to show the results to the customer, greatly saving time and the tedium of trying various clothes one by one. We believe our system can change people's traditional way of selecting clothes.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Project No. 61521002), and Research Grant of Beijing Higher Institution Engineering Research Center. This work was finished during Zhao-Heng Zheng and Hao-Tian Zhang were undergraduate students in the Department of Computer Science and Technology at Tsinghua University.

References

[1] Chen, T.; Tan, P.; Ma, L. Q.; Cheng, M. M.; Shamir, A.; Hu, S.-M. PoseShop: Human image database construction and personalized content synthesis. *IEEE Transactions on Visualization and Computer Graphics* Vol. 19, No. 5, 824–837, 2013.

[2] Zhou, S.; Fu, H.; Liu, L.; Cohen-Or, D.; Han, X. Parametric reshaping of human bodies in images. *ACM Transactions on Graphics* Vol. 29, No. 4, Article No. 126, 2010.

[3] Ferrari, V.; Marin-Jimenez, M.; Zisserman, A. Progressive search space reduction for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–8, 2008.

[4] Andriluka, M.; Roth, S.; Schiele, B. Pictorial structures revisited: People detection and articulated pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1014–1021, 2009.

[5] Johnson, S.; Everingham, M. Combining discriminative appearance and segmentation cues for articulated human pose estimation. In: Proceedings of the IEEE 12th International Conference on Computer Vision Workshops, 405–412, 2009.

[6] Sun, M.; Savarese, S. Articulated part-based model for joint object detection and pose estimation. In: Proceedings of the International Conference on Computer Vision, 723–730, 2011.

[7] Tian, Y.; Zitnick, C. L.; Narasimhan, S. G. Exploring the spatial hierarchy of mixture models for human pose estimation. In: *Computer Vision–ECCV 2012*. Fitzgibbon, A.; Lazebnik, S.; Perona, P.; Sato, Y.; Schmid, C. Eds. Springer-Verlag Berlin Heidelberg, 256–269, 2012.

[8] Yang, Y.; Ramanan, D. Articulated pose estimation with flexible mixtures-of-parts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1385–1392, 2011.

[9] Vineet, V.; Warrell, J.; Ladicky, L.; Torr, P. H. S. Human instance segmentation from video using detector-based conditional random fields. In: Proceedings of the 22nd British Machine Vision Conference, 80.1–80.11, 2011.

[10] Tompson, J. J.; Jain, A.; LeCun, Y.; Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. In: Proceedings of the Advances in Neural Information Processing Systems 27, 1799–1807, 2014.

[11] Ramakrishna, V.; Munoz, D.; Hebert, M.; Bagnell, J. A.; Sheikh, Y. Pose machines: Articulated pose estimation via inference machines. In: *Computer Vision–ECCV 2014*. Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T. Eds. Springer International Publishing Switzerland, 33–47, 2014.

[12] Carreira, J.; Agrawal, P.; Fragkiadaki, K.; Malik, J. Human pose estimation with iterative error feedback. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4733–4742, 2016.

[13] Wei, S.-E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4724–4732, 2016.

[14] Blake, A.; Rother, C.; Brown, M.; Perez, P.; Torr, P. Interactive image segmentation using an adaptive GMMRF model. In: *Computer Vision–ECCV 2004*. Pajdla, T.; Matas, J. Eds. Springer-Verlag Berlin Heidelberg, 428–441, 2004.

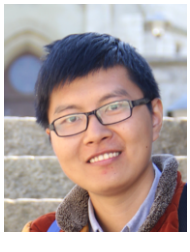
[15] Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. L. Semantic image segmentation with deep convolutional nets and fully connected CRFS. In: Proceedings of the International Conference on Learning Representations, 2015.

[16] Ladický, L.; Russell, C.; Kohli, P.; Torr, P. H. S. Associative hierarchical CRFs for object class image segmentation. In: Proceedings of the IEEE 12th International Conference on Computer Vision, 739–746, 2009.

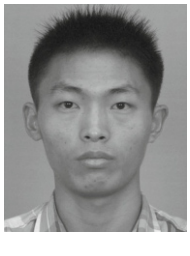
[17] Arbeláez, P.; Hariharan, B.; Gu, C.; Gupta, S.; Bourdev, L.; Malik, J. Semantic segmentation using regions and parts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3378–3385, 2012.
[18] Arbeláez, P.; Maire, M.; Fowlkes, C.; Malik, J. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 33, No. 5, 898–916, 2011.
[19] Arbeláez, P.; Pont-Tuset, J.; Barron, J. T.; Marques, F.; Malik, J. Multiscale combinatorial grouping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 328–335, 2014.
[20] Ko, B. C.; Nam, J.-Y. Object-of-interest image segmentation based on human attention and semantic region clustering. *Journal of the Optical Society of America A* Vol. 23, No. 10, 2462–2470, 2006.
[21] Cheng, M.-M.; Mitra, N. J.; Huang, X.; Torr, P. H. S; Hu, S.-M. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 37, No. 3, 569–582, 2015.
[22] Cheng, M.-M.; Warrell, J.; Lin, W.-Y.; Zheng, S.; Vineet, V.; Crook, N. Efficient salient region detection with soft image abstraction. In: Proceedings of the IEEE International Conference on Computer Vision, 1529–1536, 2013.
[23] Chen, L.-C.; Yang, Y.; Wang, J.; Xu, W.; Yuille, A. L. Attention to scale: Scale-aware semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3640–3649, 2016.
[24] Zheng, S.; Cheng, M.-M.; Warrell, J.; Sturgess, P.; Vineet, V.; Rother, C.; Torr, P. H. S. Dense semantic image segmentation with objects and attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3214–3221, 2014.
[25] Igarashi, T.; Moscovich, T.; Hughes, J. F. As-rigid-as-possible shape manipulation. *ACM Transactions on Graphics* Vol. 24, No. 3, 1134–1141, 2005.
[26] Arad, N.; Reisfeld, D. Image warping using few anchor points and radial functions. *Computer Graphics Forum* Vol. 14, No. 1, 35–46, 1995.
[27] Schaefer, S.; McPhail, T.; Warren, J. Image deformation using moving least squares. *ACM Transactions on Graphics* Vol. 25, No. 3, 533–540, 2006.
[28] Kaufmann, P.; Wang, O.; Sorkine-Hornung, A.; Sorkine-Hornung, O.; Smolic, A.; Gross, M. Finite element image warping. *Computer Graphics Forum* Vol. 32, No. 2, 31–39, 2013.
[29] Rother, C.; Kolmogorov, V.; Blake, A. "GrabCut": Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics* Vol. 23, No. 3, 309–314, 2004.
[30] Chen, T.; Cheng, M.-M.; Tan, P.; Shamir, A.; Hu, S.-M. Sketch2Photo: Internet image montage. *ACM Transactions on Graphics* Vol. 28, No. 5, Article No. 124, 2009.
[31] He, K.; Sun, J.; Tang, X. Guided image filtering. In: *Computer Vision–ECCV 2010*. Daniilidis, K.; Maragos, P.; Paragios, N. Eds. Springer-Verlag Berlin Heidelberg, 1–14, 2010.
[32] Pérez, P.; Gangnet, M.; Blake, A. Poisson image editing. *ACM Transactions on Graphics* Vol. 22, No. 3, 313–318, 2003.
[33] Shahrian, E.; Rajan, D. Weighted color and texture sample selection for image matting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 718–725, 2012.
[34] Belongie, S.; Malik, J.; Puzicha, J. Shape context: A new descriptor for shape matching and object recognition. In: Proceedings of the 13th International Conference on Neural Information Processing Systems, 798–804, 2000.
[35] Shilkrot, R.; Cohen-Or, D; Shamir, A.; Liu, L. Garment personalization via identity transfer. *IEEE Computer Graphics and Applications* Vol. 33, No. 4, 62–72, 2013.

Zhao-Heng Zheng is currently a master student at University of Michigan, Ann Arbor, USA. He received his B.S. degree from Tsinghua University in 2017. His research interests include image and video processing, semantic video understanding, and computer vision.

Hao-Tian Zhang is currently a Ph.D. student at Stanford University, USA. He received his B.S. degree from Tsinghua University in 2017. His research interests include image and video editing, and physically-based simulation.



Fang-Lue Zhang is a lecturer at Victoria University of Wellington, New Zealand. He received his doctoral degree from Tsinghua University in 2015 and bachelor degree from Zhejiang University in 2009. His research interests include image and video editing, computer vision, and computer graphics. He is a member of ACM and IEEE.



Tai-Jiang Mu is currently a postdoctoral researcher in the Department of Computer Science and Technology, Tsinghua University, where he received his Ph.D. and B.S. degrees in 2016 and 2011, respectively. His research interests include computer graphics, stereoscopic image and video processing, and stereoscopic perception.