Towards Automatic Hands Detection in Single Images

Athanasios Tsitsoulis and Nikolaos Bourbakis

Wright State University, ATR Center,
Dayton, Ohio

Abstract. Detection of hands in single, unconstrained, monocular images is a very difficult task. Localization and extraction of the hand regions, however, provides important and useful knowledge that can facilitate many other tasks, such as gesture recognition, pose estimation and action recognition. In this paper we present a simple appearance-based methodology that combines face detection and anthropometric constraints to efficiently estimate the position and regions of hands in images. It requires no training neither explicit estimation of the human pose. Experimental results illustrate the performance of the approach.

Keywords: hands detection, face detection, skin detection, segmentation.

1 Introduction

Localization and detection of human hands in unconstrained, still images is a very challenging task due several factors. The most important among them are shading, image noise, occlusions, background clutter, the high degree of human body deformability and the unrestricted positions due to in and out of the image plane rotations. Knowledge about the hands position can further benefit various tasks, such as determination of the human layout [1–4], recognition of actions from static images [5–7] and sign language recognition [8, 9].

Related literature can be roughly divided into two broad categories, methodologies that deal with detection of hands independently of the human body [10, 16–21] and those that correlate it with the pose detection problem [4, 8, 22]. Methodologies that fall into the first category produce remarkable results, but almost always rely on large training datasets in order to cover the vast space of human shapes and appearances. The most frequently used features in these methodologies are Haar-like features and HoG descriptors and the most popular classification methods employ Support Vector Machines (SVM) and boosting techniques. Methodologies that fall into the second category mostly evolve around the detection of human body parts and their overall structure, employing often the powerful graph-based pictorial structure technique. Hands detection in these cases comes as a byproduct of the very difficult and ill-posed task of body pose detection and usually assumptions and restrictions to the original problem have to be made in order to balance the trade-off between accuracy and generalization.

In this work, we present a methodology for detection of hands in single images, in the case where the upper human body is visible, upright and facing the camera and when the skin of hands is visible. The only training needed is for the initial step of the process, namely the face detection. A small training set is also required for the global skin detection process, but this process serves only as a verification step for the face detection. The rest of the methodology is mostly appearance-based and relies on the assumption that there is a connection and continuity between the (upper) human body parts. Color-based similarity combined with anthropometry provide sufficient information to support an efficient and effective hands detection methodology, without the need of training data or complex inference about the exact human body pose.

The paper is structured as follows. Section 2 provides an overview of the main components of the methodology, Section 3 describes the face detection process and Section 4 presents our anthropometric model. In Section 5 the adaptive skin detection process is discussed and in Section 6 the clothes detection process. Finally, Section 7 presents the final hands detection method and Section 8 the experimental results, while Section 9 concludes the paper.

2 Overview of the Methodology

The initial and most crucial step in our methodology is the detection of the face region, which guides the rest of the process. The information that can be extracted in this step is threefold. First, the color of the skin in a persons face can be used to match the rest of their visible skin areas. Second, the location of the face in frontal upright poses gives a strong cue about the rough location of the torso. Third, the size of the face region can further lead to the estimation of the size of body parts according to anthropometric constraints. The general scheme of the methodology can be seen in Figure 1.

Fig. 1. Main steps of the methodology

Face detection is based on OpenCV librarys implementation of the widely used Viola-Jones face detection algorithm [10]. Since face detection is the

cornerstone of our methodology, we refine the face detection results by using our skin detection and face detection algorithms. Once the elliptical region of the face is known we proceed to the foreground probability estimation. We regard as foreground the pixels with color similar to that of the face region and the pixels with color similar to a patch where the torso is expected to be with respect to the face. The skin classification is conducted using normal probability distributions, while the clothes detection employs multivariate normal probability distributions. Finally, our intuitive anthropometric model further guides and refines the hands detection process by imposing simple constraints.

3 Face Detection

Localization of the face region in our method is performed using OpenCVs implementation of Viola-Jones algorithm. Besides the obvious reasons of high performance and speed, this algorithm is based on combinations of a vast pool of Haar-like features, which essentially aim in capturing the underlying structure of a human face, regardless to the its skin color. Since skin probability in our methodology is learnt from the face region adaptively, we require an algorithm that is based on structural features.

However, Viola-Jones face detector is prone to false positive detections that can lead to unnecessary activations of our algorithm and faulty hand detections. In order to refine the results of the algorithm, we propose using the skin detection method presented in [11] and face detection algorithm presented in [12]. The skin detection method is based on color constancy and a Multilayer Perceptron Neural Network (MLPNN) trained on images collected under various illumination conditions both indoor and outdoor, and containing skin colors of different ethnic groups. The face detection method is based on facial feature detection and localization using low level image processing techniques, image segmentation and graph-based verification of the facial structure.

First, the skin probability map of the image pixels is calculated. Then, the elliptical regions of the detected faces in the image found by the Viola-Jones algorithm are evaluated according to the probabilities of the inscribed pixels. More specifically, the average skin probability of the pixels X of potential face region FR_i , for each person i , must be greater than empirical threshold $T_{GlobalSkin}$ (set to 0.7 in our experiments). If it passes the global skin test it is further evaluated by our face detector. If the facial features are detected, then FR_i is considered to be a true positive detection.

4 Anthropometric Model

The basic source of knowledge that permeates the whole hands detection methodology derives from anthropometric studies that describe the structural composition of the human body. Extraction of the face region allows us to estimate the most important structural block of the body composition, the *palm length*. In the average human body the major axis of the faces ellipse is almost equal to the

length of the palm. Let this distance be called PL from now forth. The location and size of the rest of the body components of interest can be estimated using this distance and the center of the face, $(x_{FaceCenter,i}, y_{FaceCenter,i})$.

The neck base point, approximately found at $1\ PL$ directly below the face center, can be used to define the space where the hands can be found. It is used as the center for a series of concentric ellipses that aid in better understanding the hand position and arm posture. The outer ellipse has a major axis length of $5\ PL$ and beyond this ellipse possible hand regions are automatically rejected. The zone between this ellipse and the next one, which has a major axis length of $4\ PL$, shows the zone where palms can be when the arms are fully extended. Finally, the most inner ellipse has a major axis length of $2.5\ PL$ and corresponds to the ellipse delineated by the elbows when rotated in extension. This ellipse aims in giving an indication about the degree the arm is bent.

Extending the anthropometric model to the side-view case, we show that the ellipses in the frontal view are actually ellipsoid spheres. The depth information is lost in single monocular images and it is one of the reasons that hand detection in these images is a difficult problem. In our methodology we do not explicitly deal with depth information but rather rely on body part connectivity, which can be detected in the restricted 2D space of the frontal model.

Fig. 2. Anthropometric model

5 Skin Detection

Each face region FR_i is used to construct an adaptive model for persons i skin color. We first decompose the colors of pixels that belong to FR_i into Lab and HSV color spaces. Channels L and V correspond to the intensities of the pixels and are not suitable for comparing color similarities, because they are sensitive to illumination changes. Shadows and varying illumination throughout the image might make skin pixels look different. However, we have found that a combination of the rest of the channels, namely H, S, a and b channels can result to a very suitable color space for skin detection.

First, the skin color model for each person i is estimated after fitting a normal distribution to each channel. The parameters that represent the model are

then the mean values $\mu_1, \mu_2, \mu_3, \mu_4$ and standard deviations $\sigma_1, \sigma_2, \sigma_3, \sigma_4$ for each of the H, S, a and b channels, respectively. Then, each image pixels probability of being a skin pixel is calculated separately for each channel according to a normal probability distribution with the corresponding parameters. Let $P_{1,i}(X|\mu_{1,i}, \sigma_{1,i})$, $P_{2,i}(X|\mu_{2,i}, \sigma_{2,i})$, $P_{3,i}(X|\mu_{3,i}, \sigma_{3,i})$ and $P_{4,i}(X|\mu_{4,i}, \sigma_{4,i})$ be the probabilities of image pixel X depicting skin for each of the selected channels. Then, we require that a true skin pixels should have strong probability response in all of the selected channels, so the final skin probability for each person i is:

$$P_{Skin,i}(X|\boldsymbol{\mu}_{HSab}, \boldsymbol{\sigma}_{HSab}) = \prod_{j=1}^4 P_{j,i}(X|\mu_{j,i}, \sigma_{j,i}) \quad (1)$$

The resulting skin probability image is convolved with an averaging 2D filter to smooth out outlying pixels and preserve the expected continuity of skin regions. This simple skin detection technique is able to distinguish skin from background and even skin-like regions to a good extent. In the end of this step the potential hand regions are generated.

6 Clothes Detection

Even though the adaptive skin detection results in a high number of true positives and low number of false positives, we require additional information in order to decide whether a skin region can correspond to hand region for person i . Examination and classification of each potential hand region exclusively is a very difficult problem and most probably requires extensive training. In our case, hands are regarded as skin regions connected to regions that correspond to persons clothes.

The location of the main and most visible region that represents ones clothes can be assumed to be under their face in the case of upright and frontal view. Using our anthropometric model, depicted in Figure 2, we localize and extract a patch under the shoulder line passing through the base of the neck. The width of the patch is set to 1.6 PL and its height to 2.5 PL . This area provides a good trade-off between being flexible enough to cope with small body rotations and deviations from the upright position and being able to capture an efficient number of pixels to model the clothes color probability. Here we make two assumptions. First, we assume that clothes can be sufficiently represented by small range of colors and second, we assume that the sleeves, if any, have similar color to the extracted image patch.

After the image patch is extracted, we process it using the Fuzzy Region Growing (FRG) method, published in [13], which is a computationally efficient technique that uses smoothing, edge information, homogeneity criteria and degree of dissimilarity to segment image regions. The parameters of the algorithm are set to produce a small number of regions, to avoid unnecessary segmentation. Although color homogeneity is one of our assumptions, we do not wish the model to overfit the data. Some of the resulting segments might

belong to background in case of a slightly tilted body, parts due to self occlusion or more frequently small patches of clothing, such as ties, writings, etc. that are not representative of the clothes color. Thus, we keep only the largest segment to train the clothes model and the second largest segment, if it is large enough.

For each segment $j = 1, 2$, we acquire the mean values and covariance matrices of its R, G, B components, $\mu_{R,j}$, $\mu_{G,j}$, $\mu_{B,j}$ and $C_{RGB,j}$ respectively and now calculate their multivariate normal distributions, because all channels are correlated due to the illumination component, which exists in all of them. Like in the case of skin probability estimation, the pdf responses are convoluted with an averaging filter. Then, the pdfs for each segment j are added to form the final probability:

$$P_{Clothes,i}(X|\mu_{RGB}, C_{RGB}) = \sum_{j=1}^2 P_{j,i}(X|\mu_{Rj,i}, \mu_{Gj,i}, \mu_{Bj,i}, C_{RGBj,i}) \quad (2)$$

7 Hands Detection

The final step of the methodology deals with the refinement of the potential hand regions and then the localization and extraction of the hand regions. In this step, we assume that hands are connected to the arms or clothes regions and we use this relationship to distinguish the true hand regions. In order to do so, we use a technique similar to the hysteresis thresholding technique of Canny edge detection to discover the connections. We call it double hysteresis thresholding, because it is performed in two steps. First, by using a low threshold for the skin probability, $T_{SkinLow}$, we keep the potential skin pixels and by using a high threshold for the clothes probability, $T_{ClothesHigh}$, we keep the regions that almost certainly belong to the clothes regions. Then, as in hysteresis thresholding, the potential skin regions that are connected to certain clothes regions are marked in an iterative process until convergence. Second, the same process is performed, but now using a high skin probability threshold, $T_{SkinHigh}$ and a low clothes probability threshold, $T_{ClothesLow}$. The two steps can be seen in Figures 3a and 3b. The final result is the intersection of the two steps, as seen in Figure 3c. The skin regions that have passed $T_{SkinLow}$, belong to the intersection, while being surrounded by a sufficient number of foreground pixels (clothes and skin) and being inside the outer anthropometric ellipse, are marked as hand regions. Figure 3d shows a result of successful hand region localization (the region of the face is discarded).

In the cases where the person wears clothes with long sleeves, as for instance the woman in Figure 4c, the skin region of the hands is naturally segmented, usually at palms length (PL). In the cases however where a larger portion of the arms is visible, like when the sleeves are short, further examination is needed. Due to this case, we also incorporate a skeletonization step for the skin regions, adopting the implementation in [15]. When the length of the medial axis is greater than 1 PL then the windows around its two extreme points are examined

Fig. 3. Double hysteresis thresholding, a) low probability threshold for skin and high for clothes, b) high probability threshold for skin and low for clothes, c) intersection of the previous results, d) potential hand regions

using the process discussed above. Furthermore, using the anthropometric model we can also make an assumption about the most probable hand region. More specifically, the region that is closer to the outer ellipse is more probable to represent a hand region. However, when both regions are inside the inmost ellipse inference is ambiguous, so we choose to report both regions as potential hand regions.

Finally, it should be noted that in order for the methodology to be truly detecting hand regions, the assumption that hands are visible has to be made. Classification of a region as being hand or not is difficult without further knowledge and it is beyond the scope of this work. We can use a simple technique however, mostly to differentiate arm regions from hand regions. Arm regions are in general rectangular patches and thus pose a high solidity value. Solidity here is defined as the ratio of the region area to its convex area and it reaches its maximum value 1 when the two areas completely overlap. In many cases the deformability of hands, such as in the case where fingers are open, creates a larger convex area than their original area and thus a lower value of solidity than in the case of arm regions. Using this simple measure we manage to keep the computational efficiency of our methodology. There are many techniques in the literature that specifically deal with the subject of gesture recognition that can be used instead in this point for more robust classification.

Fig. 4. Final result of our methodology, a) trimap seeds for GrowCut algorithm, b) image window encompassing the hand region, c) extracted hand regions (corresponding to the encircled face region)

8 Experimental Results

Our quantitative experiment was conducted using the sign language pose recognition dataset collected by [8]. The dataset contains frames captured from signing sequences in news broadcasts, where the background to the left of the signer is constantly changing and unknown. This is a good example of an application our methodology could be used for. Since currently our methodology is restricted to cases where the face is in frontal view and the hands do not overlap, we selected 200 images that meet these criteria to perform the testing. Table 1 presents the results of this evaluation. In our case, true positives (tp) are considered the cases where the centroid of the detected hand regions lies within a distance of less than $0.5 PL$, else they are considered false positives (fp). Finally, false negatives (fn) occur when missed hands, in other words cases where less than two hands are detected.

Table 1. Precision and recall for the sign language dataset

Precision	Recall	tp	fp	fn
93.6	94.2	374	26	23

The results show that our methodology has comparable performance with the state of the art, since the most prominent methodologies [22], [8] and [4] report 99.6%, 95.6% and 86% success rates, respectively, for the same dataset. The major cause for the false positives is cases where skin-like regions are attached to detected clothes regions, while having also size and shape that could belong to hands/arms. On the other hand, the major cause for missed detection are cases where the hands are in close proximity and/or heavy shadows exist. However, although there is clearly room for improvement, the results are encouraging, especially considering the fact that once the face area is known the methodology requires no further training for its inference.

The algorithm was also tested for several cases of the dataset in [3] are used. The dataset contains unconstrained frames from the TV show *Buffy the Vampire Slayer* and it is very challenging, because persons appear at a variety of scales, against highly cluttered background, and wear any kind of clothing. Figure 5 shows some of the successful results for this challenging dataset.

Fig. 5. Double hysteresis thresholding, a) low probability threshold for skin and high for clothes, b) high probability threshold for skin and low for clothes, c) intersection of the previous results, d) potential hand regions

9 Conclusions

This paper presents a methodology for automatic localization and detection of hands in single images. It refers to cases where the upper body parts are visible and uses an appearance-based method in combination with face detection and anthropometric constraints to guide the localization and extraction of the hand regions. It is able to produce good results in difficult cases where there is highly cluttered background and subjects appear in various scales, in a computationally efficient manner. However, as an appearance-based method it requires that there is sufficient discriminability between the foreground (skin and clothes) and the background and that there is sufficient homogeneity and continuity in the foreground regions. Future work includes proposing solutions to alleviate these problems. We also plan to use this methodology to facilitate a pose estimation algorithm, which is expected to provide a feedback mechanism that will increase the robustness of the hands detection methodology.

References

1. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: Proc. CVPR (2009)

2. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2010 (VOC 2010) Results (2010)
3. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: Proc. CVPR (2008)
4. Kumar, M.P., Torr, P.H.S., Zisserman, A.: Efficient discriminative learning of parts-based models. In: Proc. ICCV (2008)
5. Delaitre, V., Laptev, I., Sivic, J.: Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: Proc. BMVC (2010)
6. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE PAMI (2009)
7. Yao, B., Fei-Fei, L.: Grouplet: a structured image representation for recognizing human and object interactions. In: Proc. CVPR (2010)
8. Buehler, P., Everingham, M., Huttenlocher, D.P., Zisserman, A.: Long term arm and hand tracking for continuous sign language TV broadcasts. In: Proc. BMVC (2008)
9. Farhadi, A., Forsyth, D.: Aligning ASL for statistical translation using a discriminative word model. In: Proc. CVPR, pp. 1471–1476 (2006)
10. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. CVPR, pp. 511–518 (2001)
11. Kakumanu, P., Makrogiannis, S., Bryll, R., Panchanathan, S., Bourbakis, N.: Image Chromatic Adaptation using ANNs for Skin Color Adaptation. In: Proc. 16th IEEE Intl. Conf. on Tools with Artificial Intelligence, ICTAI 2004 (2004)
12. Tsitsoulis, A., Bourbakis, N.: A Methodology For Detecting Faces From Different Views. In: IEEE Int. Conf. on Tools with AI, Greece (2012)
13. Yuan, X., Goldman, D., Moghaddamzadeh, A., Bourbakis, N.: Segmentation of Color Images with Highlights and Shadows using Fuzzy-like Reasoning. Pattern Analysis and Applications 4(4), 272–282 (2001)
14. Vezhnevets, V., Konouchine, V.: GrowCut: Interactive multi-label ND image segmentation by cellular automata. In: Proc. of Graphicon, pp. 150–156 (2005)
15. Howe, N.: Skeletonization (2007) (accessed February 1, 2013)
16. Kolsch, M., Turk, M.: Robust hand detection. In: Proc. Int. Conf. Autom. Face and Gesture Recog. (2004)
17. Ong, E.J., Bowden, R.: A boosted classifier tree for hand shape detection. In: Proc. Int. Conf. Autom. Face and Gesture Recog. (2004)
18. Wu, Y., Huang, T.S.: View-independent recognition of hand postures. In: Proc. CVPR (2000)
19. Wu, Y., Liu, Q., Huang, T.S.: An adaptive self-organizing color segmentation algorithm with application to robust real-time human hand localization. In: Proc. Asian Conf. on Computer Vision (2000)
20. Zhu, X., Yang, J., Waibel, A.: Segmenting hands of arbitrary color. In: Proc. Int. Conf. Autom. Face and Gesture Recog. (2000)
21. Mittal, A., Zisserman, A., Torr, P.H.S.: Hand detection using multiple proposals. In: British Machine Vision Conference (2011)
22. Karlinsky, L., Dinerstein, M., Harari, D., Ullman, S.: The chains model for detecting parts by their context. In: Proc. CVPR (2010)