Automatic Skin Segmentation for Gesture Recognition Combining Region and Support Vector Machine Active Learning*

Junwei Han, George M. Award, Alistair Sutherland, and Hai Wu
School of Computing, Dublin City University, Ireland

Abstract

Skin segmentation is the cornerstone of many applications such as gesture recognition, face detection, and objectionable image filtering. In this paper, we attempt to address the skin segmentation problem for gesture recognition. Initially, given a gesture video sequence, a generic skin model is applied to the first couple of frames to automatically collect the training data. Then, an SVM classifier based on active learning is used to identify the skin pixels. Finally, the results are improved by incorporating region segmentation. The proposed algorithm is fully automatic and adaptive to different signers. We have tested our approach on the ECHO database. Comparing with other existing algorithms, our method could achieve better performance.

1. Introduction

As the initial step of many human body related applications such as face detection [1] and tracking [2], gesture recognition [3], objectionable image filtering [4], and videophone [13], skin segmentation has been intensively studied in the last decade. Basically, the research activities in the skin segmentation have progressed in two directions: generic colour model [1, 13] and statistical colour model [2-5, 7].

The principal idea of most existing approaches is based on the assumption that skin colour is quite different from colours of other objects and its distribution might form a cluster in some specific colour-spaces. The generic colour model defines a fixed colour range to separate skin from non-skin pixels. Generally, the fixed colour range is achieved empirically using some collected training instances. Apparently, this group of methods is easy to implement and adequate for real-time systems. However, they cannot handle illumination and human skin variations. To reduce the limitations, the so-called statistical colour model is proposed to determine the skin pixels. It considers skin segmentation as a typical one-class or two-class classification problem. Firstly, the skin colour distribution is estimated by a Gaussian model or the histogram based technique using a large number of training data. Then, a Bayesian classifier or other learning algorithms is applied to classify the skin and non-skin pixels. The classifiers are constructed adaptively. Therefore, they allow the segmentation to be robust to the different lighting conditions and human skin types. The drawbacks of the statistical model are twofold. The former is that it often works well under the premise that the skin colour distribution is a single or mixture Gaussian. Unfortunately, the assumption is not satisfied in many cases, which might somewhat influence the segmentation accuracy. The other disadvantage is that it requires a large amount of training data that cover different human skin appearances. It is very expensive to collect so many training instances.

In gesture recognition systems, the skin segmentation plays the role of automatically detecting the hands and face from the given videos. Compared with other applications, videos for gesture recognition systems generally contain very few signers and the signers' skin colours keep consistent across the frames. It means we do not need to get many samples to train the model. Moreover, the useful information from previous frames could be adopted to process the current frame. Hence, from the viewpoint of collecting training data, one effective idea is to use the pixels from the first couple of frames as the training samples.

Based on the specific characteristics of the gesture recognition system, this paper introduces a novel skin segmentation approach integrating region and SVM active learning. The framework consists of two stages: training stage and segmentation stage. In the training stage, first, for the given gesture video, a generic skin colour model is applied to the first several frames, which obtains the initial skin areas. Afterwards, a binary classifier based on SVM active learning is trained using obtained initial skin areas as the training set. In the segmentation stage, the SVM classifier is incorporated with the region information to produce the final segmentation results. Fig. 1 displays the basic architecture of the proposed work.

The first author is supported by the EU Marie Curie Incoming International Fellowship, project 509477.

Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR'06)
0-7695-2503-2/06 $20.00 © 2006 IEEE

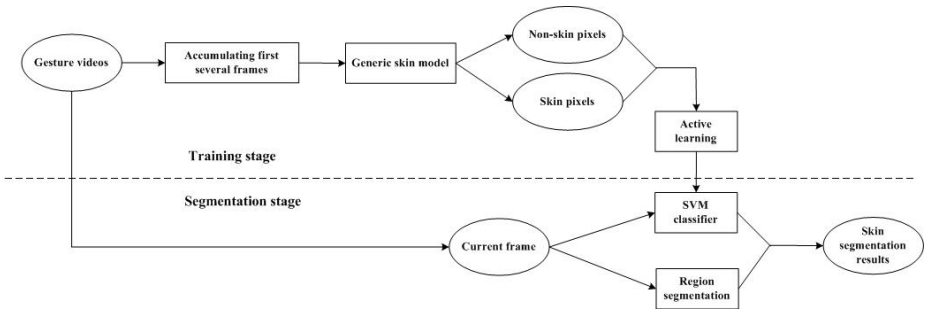

Fig.1 The basic architecture of the proposed work.

We summarize the contributions of our work below.

1. Using region information. Most traditional skin segmentation is based only on pixels. In contrast to the individual pixel, region information can reduce the noise and illumination variation. Our experiments demonstrate the combination of region and SVM can improve the segmentation results.

2. Applying the SVM active learning to skin segmentation. As analyzed before, in gesture recognition system, skin segmentation does not need a large-scale training set. Thus, SVM is very suitable due to its outstanding capability in dealing with classification with small sample size. Also, active learning is used to select the most informative training subset for SVM, which leads to the fast convergence and better performance.

The rest of this paper is organized as follows. In Section 2, we report the proposed work. In Section 3, we show the experimental results on the ECHO database. Additionally, the comprehensive comparisons of our methods with other traditional methods are provided. Finally, conclusions are drawn in Section 4.

2. The proposed approach

2.1 The generic skin model

In our work, the target of adopting the generic skin model is to collect some training data for SVM classifier. The generic skin model is implemented by defining a fixed skin colour range in one colour-space. Here, like [1], we represent the skin pixels in the RGB space as follows:

$$\begin{aligned}&\text{(i) under uniform daylight}\\&(R>95)\,\&\,(G>40)\,\&\,(B>20)\,\&\,(\max\{R,G,B\}\\&-\min\{R,G,B\}>15)\,\&\,(|R-G|>15)\,\&\,(R>G)\\&\&\,(R>B)\,\&\,(G>B)\qquad(1)\\&\text{(ii) under the flashlight}\\&(R>220)\,\&\,(G>210)\,\&\,(B>170)\,\&\,(|R-G|\le 15)\\&\&\,(R>B)\,\&\,(B>B)\,\&\,(G>B)\end{aligned}$$

2.2 SVM active learning for skin segmentation

SVMs are invented by Vapnik [8]. Concerning SVM in a binary classification setting, given linear separable training set $\{\mathbf{x_1},\mathbf{x}_2,....,\mathbf{x}_n\}$ with their labels $\{y_1,y_2,....,y_n\}, y_i\in\{-1,1\}$, the SVM is trained and optimal hyperplane is yielded, which separates the training data by a maximal margin. Specifically, the optimal hyperplane might be found by solving an optimization issue:

$$\begin{aligned}&\text{Minimize}:\phi(\mathbf{w})=\frac{1}{2}\|\mathbf{w}\|^2\\&\text{Subject to}: y_i(\mathbf{w}\bullet\mathbf{x}_i+b)\ge 1\end{aligned}\qquad(2)$$

Here

$$\mathbf{w}=\sum a_i y_i \mathbf{x}_i \qquad(3)$$

The optimal hyperplane divides the data points into two sides. Points lying on one side are labelled -1, and the other points are marked 1. When a new example is inputted for classification, a label (1 or -1) is issued by its position to the hyperplane that is:

$$f(\mathbf{x})=sign(\textstyle\sum_i \alpha_i y_i(\mathbf{x}_i\bullet\mathbf{x})+b) \qquad(4)$$

For the case of data being not linearly separable, the SVM first projects the original data to a higher dimensional space by a Mercer kernel function K, and then linearly separates them. The corresponding nonlinear decision boundary is:

$$f(\mathbf{x})=sign(\textstyle\sum_i \alpha_i y_i K(\mathbf{x}_i\bullet\mathbf{x})+b) \qquad(5)$$

Technically, the kernel functions could be Gaussian RBF kernels, polynomial kernels, and Laplacian RBF kernels.

The SVM can be easily to apply to the skin colour segmentation. Given a gesture video, the generic skin model is performed on the first several frames so that the training set containing skin and non-skin data could be obtained. Afterwards, the SVM classifier is constructed using the training set from the previous frames to segment the future frames one by one.

In practice, one problem is the imbalance of the training data. That is, the number of the negative examples (non-skin pixels) is far larger than the number of the positive examples (skin pixels). Fig. 2 displays

the fact. The left picture is the original image, and the right one is the segmented results. In the right image, the points labelled by green colour are the skin pixels and other points mean the non-skin pixels. The imbalance of training examples may make the learning less reliable. Moreover, it results in long learning time. A feasible way to reduce the limitation is to use active learning. Active learning is named in contrast to the traditional passive learning. Most machine learning approaches belong to passive learning because they are usually based on the entire training set or randomly selected data [11]. In contrast, active learning tries to find the most informative data to train the classifier. Its goal is to achieve better performance and faster convergence with less training examples. Lately, active learning has been successfully introduced to document classification [10], image retrieval [11], and text classification [12]. Whereas, to our best knowledge, very little work has been done in the skin segmentation field.

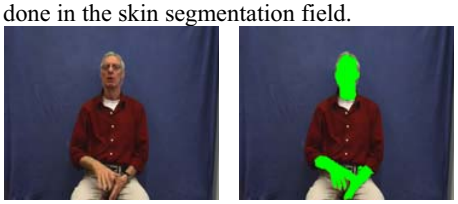

Fig. 2 The imbalance of the training examples.

The key idea of active learning is to extract the most informative samples from all available training data. As for our application, we attempt to find the small but informative subset of negative examples with a similar size to the training set of positive examples. According to the principle of maximally reducing the SVM version space [11, 12], the importance of one instance point depends on its distance to the dividing hyperplane. The instances closer to the SVM hyperplane generally give the larger influence to the learning so that they are more informative than other instances. This motivates us to design a similarity-based sampling strategy to select more informative negative examples for our specific application.

Let F be the training set. F^+ and F^- are positive (skin pixel) and negative example (non-skin pixel) set, respectively. $F = F^+ \bigcup F^- \;\&\; F^+ \bigcap F^- = \varnothing$. We hope to get the small subset of negative examples F^-_{Active} by the active learning. Here, $F^-_{Active} \subset F^-$. First, a region segmentation scheme JSEG [9] is employed to segment F^- into different regions, $R_1^{F^-}, R_2^{F^-}, ..., R_N^{F^-}$. $\sum_{i=1}^{N} R_i^{F^-} = F^-$. Second, the similarities between each $R_i^{F^-}$ and F^+ are described by the colour histogram based distances. More specifically,

$$D(R_i^{F^-}, F^+) = \| H(R_i^{F^-}) - H(F^+) \| \quad (6)$$

where $H(\bullet)$ is the colour histogram vector. The smaller distance between $R_i^{F^-}$ and F^+ indicates they are more similar. In the feature space, the skin pixels F^+ generally could form a cluster. If one negative instance is closer to F^+, it is closer to the SVM hyperplane as well. Therefore, it is more likely to be the informative examples. Finally, we sample the negative examples according to a principle called "most similar highest priority". To be specific, more negative instances are extracted from the $R_i^{F^-}$ with the smaller distance to F^+, but less negative instances are selected from the $R_j^{F^-}$ with the larger distance to F^+. The sampled examples construct the F^-_{Active}, and its size is approximately same to the size of F^+. Apparently, the advantage of our similarity-based sampling strategy is that not only can it get more informative examples, but also the obtained set F^-_{Active} covers all kinds of negative examples from different regions.

In summary, the SVM active learning for skin segmentation is fulfilled by the following steps:

1. Apply the generic skin model to the first several frames to obtain F^+ and F^-;
2. Segment F^- into different regions $R_1^{F^-}, R_2^{F^-}, ..., R_N^{F^-}$, and compute the distances between each $R_i^{F^-}$ and F^+ in the colour feature space;
3. Construct the F^-_{Active} from F^- in accordance with similarity based sampling scheme;
4. Train the binary SVM classifier using F^+ and F^-_{Active};
5. Classify every current frame into skin and non-skin pixels by the trained SVM.

2.3 Combination of SVM active learning and region information

Although the performance of SVM active learning is outstanding, it cannot produce perfect skin segmentation results due to noise and illumination variation. However, region information is considerably robust to noise and illumination variation. Hence, in order to solve this problem, this paper incorporates region information to further refine the segmentation result. First, the JESG algorithm [9] is adopted to parse the

frame into regions. Then, if the majority of points of one region belong to skin pixels, the whole region is declared as the skin area. To be exact, one region R_i satisfying

$$\frac{NS(R_i)}{NT(R_i)} > \eta \quad (7)$$

is decided as skin area. Here, $NS(R_i)$ denotes the number of the skin pixels in the region R_i, $NT(R_i)$ refers to the number of pixels in R_i, and η is an empirically defined constant.

3. Experimental results

We tested the proposed work with 8 gesture video sequences from the ECHO* database. They were captured with different signers and under different lighting conditions. Almost every video sequence is over 15 minutes long. To quantitatively evaluate our work, we randomly picked 240 frames from 8 video sequences, then two students were invited to manually segment skin pixels to construct the ground truth. All algorithms in the experiments were implemented by Matlab, and the classifiers were trained using accumulated three frames. As in [5], three metrics, correct detection rate (*CDR*), false detection rate (*FDR*), and overall classification rate (*CR*) were employed to measure the performance of the techniques. They are described as follows [5]:

$$\begin{aligned} &CDR: \text{percent of correctly classified skin pixels;} \\ &FDR: \text{percent of wrongly classified non-skin pixels;} \\ &CR: \quad \frac{N_S}{\max(N_S^A, N_S^G)} \times 100\% \end{aligned} \quad (8)$$

where N_S is the number of skin pixels detected both by the algorithm and the ground truth, N_S^A is the number of skin pixels detected by the algorithm, and N_S^G is the number of skin pixels detected by the ground truth.

Three aspects of experiments were constructed to evaluate the proposed algorithm. In Section 3.1, we test the performance of active learning. In Section 3.2, we examine the capability of combining region information. Finally, comparisons with some traditional skin segmentation techniques are reported in Section 3.3.

3.1 Performance test of the active learning

To verify this performance of SVM active learning, we compared the SVM classifiers with and without active learning using our test data. Fig. 3 shows one set of sample results. The first, second and third image display the original frame, the SVM without active learning, and SVM with active learning, respectively. Table 1 lists the statistical results including the precision and training time. As can be seen from the experimental results, the SVM active learning is superior in both accuracy and computational complexity. It can enhance the overall accuracy almost by 6%, and decrease average training time by 114 seconds.

3.2 Evaluation of combining region information

This experiment is to evaluate the segmentation results with and without region information. Fig. 4 displays some sample results, and Table 2 lists the statistical precision comparisons. In Fig. 4, the first column shows the original frames, the second column shows the segmentation results without region information, and the third column shows the results with region information. Clearly, the algorithm with region information is better, which can reduce the noise and refine the segmentation results. Incorporation of region information enhanced the overall accuracy by 9%.

3.3 Comparisons with traditional skin segmentation techniques

To demonstrate the effectiveness of the proposed work, we compared with two existing skin segmentation algorithms, generic skin model [1, 13] and Gaussian model [3-5]. The Gaussian models [5] can be described as follows. They employed the Bayesian decision rule

$$\frac{p(c \mid skin)}{p(c \mid nonskin)} \geq \xi \quad (9)$$

to classify the skin and non-skin pixels. Here, $p(c \mid skin)$ and $p(c \mid nonskin)$ refer to the probability density function (pdf) of skin and non-skin colour, respectively. ξ is a threshold. The colour pdf could be modelled as a single Gaussian:

$$G(c) = 2\pi^{-1} |\sum|^{1/2} e^{-\frac{1}{2}(c-\mu)^T \sum^{-1}(c-\mu)} \quad (10)$$

or Gaussian mixture:

$$\begin{aligned} &GM(c) = \sum_{i=1}^{k} \omega_i G_i(c) \\ &\sum_{i=1}^{k} \omega_i = 1 \end{aligned} \quad (11)$$

where μ is the mean vector, $\sum$ is the covariance matrix, and k is the number of the mixture components, respectively. In [5], Phung proposed two strategies: only modeling skin pixels as Gaussian (called one-Gaussian in this paper) and modeling both skin and non-skin pixels as Gaussian (called two-Gaussian in this paper). In this experiment, we implemented these two strategies. Notice we used Gaussian mixture to model pdfs. Fig. 5

* ECHO is a sign language database in Europe. It is available on http://www.let.ru.nl/sign-lang/echo/.

shows some results. The segmentation results by generic model, one-Gaussian, two-Gaussian, and the proposed approach are displayed in the first, second, third, and fourth column, respectively. Table 3 lists the statistical accuracy comparisons. As we can see from the comparison results, the proposed model has the highest overall accuracy with the second lowest false detection rate. Although the two-Gaussian model has the best correct detection rate, its false detection rate is worst.

4. Conclusions

In this paper, an adaptive skin segmentation algorithm for gesture recognition system has been proposed. A binary SVM classifier was trained using the training data automatically collected from the first several video frames. More importantly, the active learning and region segmentation were combined to further improve the performance. One important advantage of the proposed work is it is easy to implement and does not need human labour to construct the training set. In addition, it may be efficiently incorporated in a gesture recognition system or other human body related applications with minor revision. A set of comprehensive experiments tested on real-world sign language videos demonstrated the proposed work is promising.

5. References

[1] J. Kovac, P. Peer, and F. Solina, "Human Skin Color Clustering for Face Detection," in *Proc. of EUROCON 2003*, Finland, 2003, pp. 144 – 148.

[2] J. Yang, W. Lu, and A. Waibel, "Skin-color Modeling and Adaptation," in *Proc. of Asia Conf. Computer Vision*, Hong Kong, 1998, pp. 687-694.

[3] X. Zhu, J. Yang, and A. Waibel, "Segmenting Hands of Arbitrary Color," in *Proc. of IEEE Conf. Automatic Face and Gesture Recognition*, France, 2000, pp. 687-694.

[4] Q. Zhu, C. T. Wu, K. T. Cheng, and Y. L. Wu, "An Adaptive Skin Model and Its Application to Objectionable Image Filtering," in *Proc. of ACM Multimedia*, USA, 2004, pp. 56-63.

[5] S. L. Phung, A. Bouzerdoum, and D. Chai, "Skin Segmentation Using Color Pixel Classification: Analysis and Comparison," " *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 21, Jan. 2005, pp. 148-154.

[6] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A Survey on Pixel-based Skin Color Detection Techniques," in *Proc. Of Graphicon-2003*, Russia, 2003, pp. 85-92.

[7] M. J. Jones and J. M. Rehg, "Statistical Color Models with Application to Skin Detection," *Int'l Journal of Computer Vision*, Vol. 46, 2002, pp. 81-96.

[8] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

[9] Y. Deng and B. S. Manjunath, "Unsupervised Segmentation of Color-texture Regions in Images and Video," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, Aug. 2001, pp. 800-810.

[10] G. Schohn and D. Cohn, "Less is More: Active Learning with Support Vector Machines," in *Proc. Of Int'l Conf. on Machine Learning*, USA, 2000, pp. 839-846.

[11] S. Tong and E. Chang, "Support Vector Machine Active Learning for Image Retrieval," in *Proc. of ACM Multimedia*, Canada, 2001, pp. 107-118.

[12] S. Tong and D. Koller, "Support Vector Machine Active Learning with Applications to Text Classification," *Journal of Machine Learning Research*, 2001, pp. 45-66.

[13] D. Chai and K. N. Ngan, "Face Segmentation Using Skin-color Map in Videophone Applications," *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 9, Jun. 1999, pp. 551-564.

Fig. 3 Experimental samples with and without active learning.

Table 1. The Statistical Precision and Training Time Comparisons with and without Active Learning

	CDR (%)	FDR (%)	CR (%)	Training time (s)
SVM without active learning	85.12	2.43	61.97	121.14
SVM with active learning	82.83	1.39	67.60	7.33

Fig. 4 Comparison results with and without region information.

Table 2 The Statistical Precision Comparisons with and without Region Information

	CDR (%)	FDR (%)	CR (%)
The proposed method without region	82.83	1.39	67.60
The proposed method with region	86.34	0.96	76.77

Fig. 5 Sample results of generic skin model, one-Gaussian model, two-Gaussian model, and the proposed model.

Table 3 The Statistical Accuracy Comparisons of Existing Models and the Proposed Model

	CDR (%)	FDR (%)	CR (%)
The generic skin model	71.51	0.79	65.10
One-Gaussian model	72.74	1.04	66.85
Two-Gaussian model	90.88	4.41	57.06
The proposed model	86.34	0.96	76.77