Finding Relevant Image Content for mobile Sign Language Recognition

SUAT AKYOL, PABLO ALVARADO
Department of Technical Computer Science
RWTH Aachen, Germany
{akyol, alvarado}@techinfo.rwth-aachen.de

ABSTRACT

We are currently developing a vision-based sign language recognition system for mobile use. This requires operability in different environments with a large range of possible users, ideally under arbitrary conditions. In this paper, the problem of finding relevant information in single-view image sequences is tackled. We discuss some issues in low level image cues and present an approach for the fast detection of a signing persons hands. This is achieved by using a modified generic skin color model combined with pixel level motion information, which is obtained from motion history images. The approach is demonstrated with a watershed segmentation algorithm.

KEY WORDS

Image Processing, Low-level Image cues, Segmentation, Sign Language Recognition

1. Introduction

So far deaf people have almost completely been excluded from technical innovations that have been important for social life. Examples are media (radio, tv) and mobile communication. Now, an EC founded project called "WISDOM" has the goal to create a mobile communication device for deaf people, which shall allow them to benefit from European 3rd generation telecommunication networks. Our task within the project is to develop a vision-based sign language recognition system for this device, in order to provide access to future services like sign language translation. The required properties of this recognition system are:

- single view image acquisition
- operability despite unknown user / environment
- near real-time performance

Considering the manual parameters (posture, pose, position and motion of the hands) as input, the pattern matching problem has already been solved with data-gloves [1, 2] and imaging devices in known environments [3, 4, 5]. Imaging devices are more reasonable for practical applications, since they are less cumbersome and allow also to capture facial expression. Finding and extracting information about human hands and face from image sequences is also interesting for a whole range of other applications [6, 7], hence there is a lot of research on this area. The problem of operation in environments other than a laboratory is an active research topic. We want to address it with regard to the detection of hands for vision based sign language recognition.

We start with an introduction to existing approaches that are relevant in the context of our application and explain our concept of a processing scheme using multiple low level image cues. Then details are given about our color model, which is an adaptation of a generic skin-color model, afterwards the motion extraction method is explained. We present some results with a segmentation algorithm based on watersheds, although this is only a preliminary example. At last, some unsolved problems are discussed, which point out to remaining future work.

2. Image Processing for Sign Language Recognition

An early vision based system for the recognition of sign language is presented in [3]. Starner and Pentland use a single video camera and uniformly colored gloves to aid the segmentation and the feature extraction processes. Later they also show, that a user-calibrated skin color model delivers similar results in a known environment.

Hienz et. al [5] use color coded gloves which allow to obtain detailed information about each finger of the dominant hand. The environment is restricted to an empty white background. In arbitrary environments however, neither skin color nor any other color can be guaranteed to appear only within the object of interest, which is the hand. Thus, relying on color information only is not sufficient, not even with the aid of colored gloves.

The usual extension for higher reliability is the combination of multiple image cues. The term "image cues" denotes information, that can be extracted without higher semantic information about the actual image content. Motion is considered as a good supplement for color, since gestures are dynamic acts.

Therefore Imagawa et al. [8] use a subsequent integration scheme, first using color and then motion information. They apply histogram backprojection to an image and segment it into connected regions. These regions are tracked by a Kalman filter to find unique correspondences for hands and face. The effect of large skin colored spots in the background is not considered, but it might be a problem, since the segmentation is

sequential and thus strongly influenced by the color information.

Yang and Ahuja [9] do it in reverse order, they use motion and then color. First, several steps are performed for motion segmentation, then the extracted regions are scored for skin color likeliness and finally adjacent regions are merged until the shape is elliptic or rectangular. These regions are assumed to be the hands and the face. However, the authors only present results for a person standing in front of a uniform background. Besides, the computational cost of this approach is relatively high.

Both approaches integrate image cues sequentially into image analysis. Our idea is to combine image cues simultaneously into a single probability map, in order to indicate relevant content for subsequent processing steps. Extension for additional cues is straight forward this way and thus very flexible. The basic idea behind this has been used in visual attention approaches for indicating salient image content (e.g. [10]).

Clearly, the most information is always contained in the original image. Any crucial processing like segmentation or tracking should therefore be done on the original image, and the map should be considered as an additional aid. The following image is a graphical representation of this concept, with the target of image segmentation.

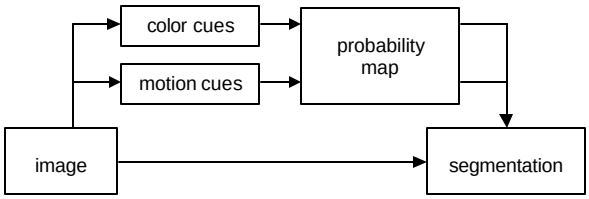

Figure 1. Image processing scheme. Simultaneous and supporting integration of image cues is proposed for segmentation instead of sequential processing.

3. Generation of Probability Map

This chapter gives details about the generation of the probability map from color and motion cues.

3.1 Color Cues

In different works it is stated, that human skin color is similar in hue across all races but differs in intensity. For this reason skin color is often modeled as probability distribution in the chromaticity plane, while intensity information is omitted. This works quite good if user and environment are known [11, 12], but in natural scenes hue and saturation can be influenced by color and intensity of unknown light sources. Also, omitting intensity reduces the three dimensional color space to two dimensions and thus loses important information.

Jones and Rehg [13] propose a generic skin color model to overcome this problem. They create histograms in RGB-space for skin and non-skin color distribution on the basis of more than 18.000 images (nearly 2 billion pixels). These were collected from a web-search and include a large variety of illumination conditions. Half of the pictures contain skin and have been segmented by hand to obtain the skin regions, the other half contains no skin. The histograms are used to calculate skin color probability for a single RGB-colored pixel with Bayes' theorem (see equation 1), which transforms an image into a probability map.

$$p(s \mid rgb) = \frac{p(rgb \mid s) \cdot p(s)}{p(rgb \mid s) \cdot p(s) + p(rgb \mid \neg s) \cdot p(\neg s)} \qquad (1)$$

Where $p(s)$ and $p(\neg s)$ is the a priori probability for observing a skin or non-skin pixel, that was calculated for the given training set, and with $p(s) + p(\neg s) = 1$. $p(rgb|...)$ denotes the value of the respective histogram bin at the coordinates r,g,b.

Jones and Rehg use simple thresholding to classify a color into skin or non-skin class and obtain a best result of 90 % correct versus 14 % false detections. Other work [14, 15] has proven that using color spaces other than rgb doesn't improve the performance, since discriminability is determined by the differences of skin and non-skin entries in color space, which get transformed into other color spaces, too.

For our purpose it is not appropriate to calculate the a priori probabilities as described by Jones, because our application is biased towards sign language recognition. Since we don't have substantial comparable training material, we apply a region adaptive method for computing them from the given image.

Starting with initial values of $p(s) = p(\neg s) = 0.5$, we set up the probability map like Jones proposed it. In the next step the value at the image coordinate (x,y) is regarded as the position-dependent a priori probability $p(s|x,y)$. The new value, that a pixel of a given color is skin, is computed as the mean value of an 8-neighborhood.

$$p(s \mid rgb, x, y) = \frac{1}{9} \sum_{x=x-1}^{x+1} \sum_{h=y-1}^{y+1} \frac{p(rgb, x, y \mid s) \cdot p(s \mid x, h)}{p(rgb, x, y \mid s) \cdot p(s \mid x, h) + p(rgb, x, y \mid \neg s) \cdot p(\neg s \mid x, h)} \qquad (2)$$

This scheme can be used repeatedly, but a visual improvement is already noticeable after one iteration. The result with an example frame from the mpeg test sequence "silent" can be seen in figure 2. The original image is not included in the color model, nevertheless interesting regions get high probabilities. Unfortunately there are irrelevant areas in the background, that get high skin color probabilities, too. Color alone is obviously not sufficient here.

(a) Original (b) Non-adaptive method (c) Adaptive method

Figure 2. Result of skin color model applied to one frame of the mpeg test sequence "silent". Skin probability maps are shown for $p(s) = p(\emptyset\ s) = 0.5$ in (b) and for the region adaptive method for estimating a priori probabilities in (c). Note that contrast is enhanced and pixel noise is reduced.

3.2 Motion Cues

A whole research area is concerned with pixel level motion detection, known as optical flow computation. It summarizes all methods for estimating the direction and magnitude that each single pixel of an image moved between successive frames. The method of Horn & Schunk is one of the first algorithms for optical flow estimation and still amongst the best performing. Yet, the method of Lukas & Kanade is reported to be the best with regard to processing speed and accuracy [16]. But optical flow techniques generally rely on intensity information, i.e. they can not detect motion at borders with different color, if intensity is equal. Besides, they are associated with high computational load.

If only the motion magnitude is searched for, the much simpler motion history images (MHI) can be regarded as an approximation [17]. MHIs are in principle images, where the decaying difference of subsequent frames is overlaid over each other.

$$H(x,y,t)=\begin{cases} 1.0 & \text{if } D(x,y,t)=1 \\ \max(0.0, H(x,y,t-1)-\frac{1}{t}) & \text{else} \end{cases} \quad (3)$$

$H(x,y,t)$ is the motion history image at time index t and t is the decaying time constant. A value of 1 in $D(x,y,t)$ indicates that the difference in either channel of red, green or blue between successive images is larger than a defined threshold and thereby takes color changes into consideration. The choice of threshold corresponds to a minimum motion level, that must be observable. We define the threshold as a fixed proportion of the mean difference between two images.

MHIs can be computed very fast. Unfortunately they tend to amplify the motion at the boundaries of an object and produce sharp steps. This can be reduced by convolving it with a gaussian kernel and results like depicted in figure 3 can be obtained, where the intensity level can be regarded as probability for motion.

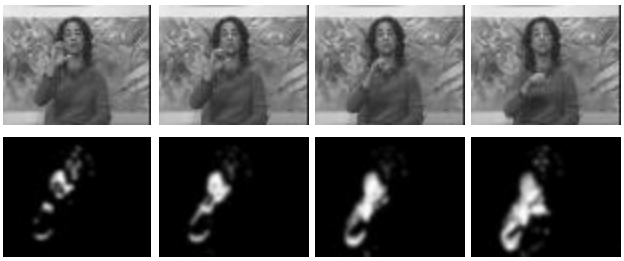

Figure 3. Top row shows four subsequent frames of the sequence. Bottom row is the motion history image after gaussian filtering.

3.3 Combination of Cues

Under the assumption that color and motion cues produce stochastically independent probability maps, the correct method for computing one combined map is pixel wise multiplication. The resulting probability map can be seen in figure 4 for some example frames. It should again be pointed out, that these results were obtained without knowledge about image content and without any kind of calibration. Yet, the visual impression for these example frames is good. The processing speed on a 500 Mhz personal computer with QCIF resolution is about 25 fps, including the filtering part of the MHIs. QCIF is a common resolution for video conferencing applications.

However, the probability map represents vague information, so it shouldn't be expected, that the map can provide good segmentation results directly, for example by thresholding. First of all, borders are unlikely to match the edges of the actual object. And second, the map should preferably be generated from downsampled images to be less noisy, which means that the resolution might not be detailed enough for accurate segmentation. Thus the next section is an example for segmentation according to the scheme in figure 1.

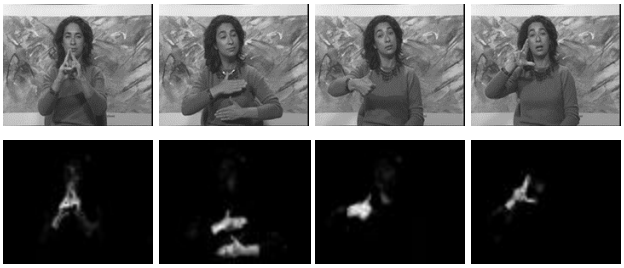

Figure 4. Result of combining color and motion into one probability map. Intensity corresponds to probability.

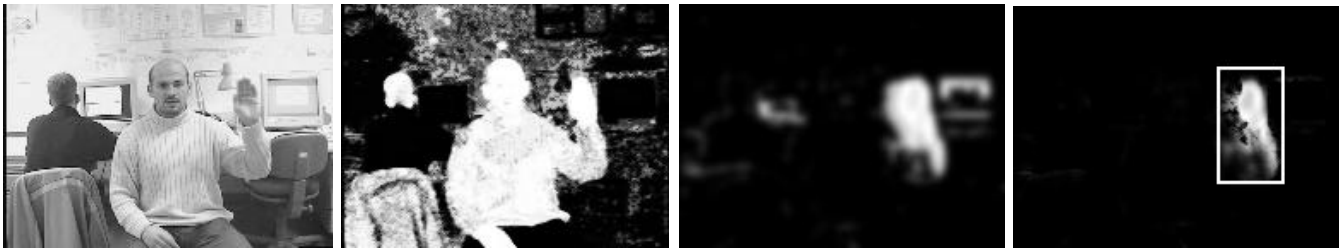

Figure 5. The leftmost image is a frame of a waving sequence, taken in our laboratory. The other images are from left to right: skin color map, motion map, combined map with bounding box of relevant region.

4. Using Image Cues for Segmentation

Figure 5 shows our laboratory and the probability maps for color, motion and the combination of both. The color map yields high probabilities for the left persons head, and the clothing on the left chair, as well as the signers clothing. The motion map captures flickering on both computer monitors. When combined, all these distortions are eliminated, leaving a representation of the waving hand and a little piece of the arm. Again, processing is done without any additional knowledge about the environment and without special illumination.

We now address the problem of region segmentation in this case as an example for how to apply the probability map. Here the fast watershed algorithm as described in [18] is used. It still is time consuming, i.e. about 2.5 seconds on a size of 768×576 with the previously described hardware. Furthermore, watersheds tend to deliver heavily oversegmented results, which increases even more with additional noise. Prior noise filtering can reduce oversegmentation in return to undesired loss of border precision. Applying the image cues can help in two ways. First, speedup can be achieved by restricting watershed search to relevant sub-regions, and second, oversegmented regions can be grouped to yield large connected regions with high border precision. It is sufficient to use the lower QCIF resolution for probability maps, in order to lessen computational load.

The areas of interest can be extracted from the map by thresholding, searching for connected regions and using a slightly enlarged bounding box. For the given example there is only one, which represents an area of approximately 160×280 pixel in the original image. Watersheds can be computed on this area in only 250 msec. Let $p(x,y)$ be the probability map, Q a predefined threshold, $w(x,y)$ the watershed region mask and $s(x,y)$ the segmentation result. Then the following grouping algorithm can be used to perform the segmentation.

```
1. Initialize s(x,y) with 0
2. Scan p until p(x,y) ≥ Θ
3. For corresponding region A in w
   3.1 compute mean probability
   3.2 count number of pixel N with
       p(x,y) ≥ Θ
4. If mean ≥ Θ and N ≥ 50 %, then
   label A in s
5. Delete A from p
6. Repeat with 2. until p is scanned
   completely
```

The result can be seen in figure 6. The hand segmentation is very accurate at borders, but also captures a fraction of the arm, due to the skin color similarity of the users sweater (see figure 5.).

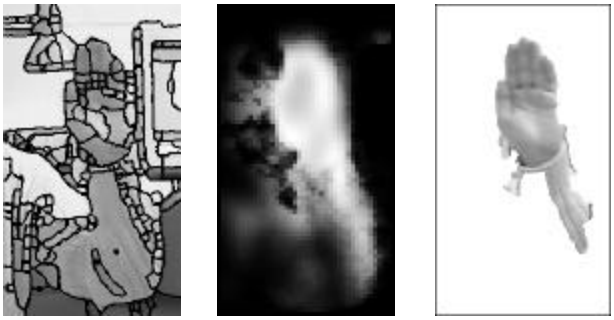

Figure 6. Left: Watersheds overlaid over original. Middle: Probability map. Right: Segmentation result.

The drawback of this procedure is, that the absence of motion (e.g. the user takes a short rest) leads to a total loss of information in the probability map. Tracking is required to stay focused on the corresponding region. the described method is also not capable of separating two hands when they overlap each other. Borders between hands might still be accurate in the watershed image, but more sophisticated methods must be used to assign adjacent regions to either hand. Likewise, moving the hand in front of the face is a problem, since the face can move, too. An active shape tracker [19] might be a solution. Another problem are skin colored items in the background, like a wooden cupboard. A runtime adaption of the skin color model is required, in order to restrict the very general skin color model to fit the current user. The user's face could be suitable for automatic extraction of skin color, since faces have salient texture and rigid shape and therefore can be detected more reliably [7]. All these drawbacks show, that some tracking/prediction and a user/hand model are necessary, in order to cope with

overlapping and occlusion. Nevertheless, the proposed method of image cue utilization can be a help in any case.

5. Summary and Outlook

This work presents an approach for the fast detection of gesturing hands from image sequences. It is based on the combination of the low level image cues of color and motion. The color map is a generic skin color model, which is extended for region adaptive estimation of a priori probabilities. Motion probabilities are obtained from motion history images, also known as temporal templates. Combination of both yields a powerful method for detecting interesting image content for vision-based sign language recognition. The required processing power of tasks like region segmentation can be reduced significantly this way.

The current work is merely a first step towards environment and user independent sign language recognition. Extraction of additional image cues is one future topic. Texture and edges are interesting, because hands don't have salient texture and are limited by borders. Future work will also deal with adapting the color model to the user at runtime. Also, tracking and prediction techniques will be applied, in order to deal with occlusion and overlapping. Then, appearance and model based feature extraction will be used, leading to hand and user modeling.

Acknowledgement

This work has been done in the scope of the EC-founded Project "WISDOM- Wireless Information Services for Deaf People on the Move".

References

[1] A. Braffort, A Gesture Recognition Architecture for Sign Language, The Second Annual ACM Conference on Assistive Technologies, Vancouver, Canada, 1996, 102-109.

[2] R.H. Liang, M. Ouhyoung, A Real-Time Continuous gesture Recognition System for Sign Language, IEEE Third International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, 558-565.

[3] T. Starner, J. Weaver, A. Pentland, Real-Time American Sign Language Recognition from Video Using Hidden Markov Modells, M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 375, 1996.

[4] C. Vogler, D. Metaxas, Adapting Hidden Markov Models for ASL recognition by using three-dimensional computer vision methods, IEEE Intl. Conf. on Systems, Man and Cybernetics, Orlando, USA, 1997, 156-161.

[5] H. Hienz, K. Grobel, Automatic Estimation of Body Regions from Video Images, in I. Wachsmuth and M. Fröhlich (Eds.) *Gesture and Sign Language in Human-Computer Interaction* (Berlin: Springer-Verlag,1998), 135-145.

[6] V. Pavlovic, R. Sharma, T. Huang, Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7), 1997, 677-695.

[7] M. Pantic, L.J.M. Rothkrantz, Automatic Analysis of Facial Expressions: The State of The Art, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12), 2000 1424-1445.

[8] K. Imagawa, S. Lu, S. Igi, Color-Based Hands Tracking System for Sign Language Recognition, IEEE Third International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, 462-467.

[9] M.H. Yang, N. Ahuja, Extracting Gestural Motion Trajectories, IEEE Third International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, 10-15.

[10] L. Itti, C. Koch, E. Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, IEEE Trans. Pattern Analysis and Machine Intelligence, 20(11), 1998, 1254-1259.

[11] M. Störring, H.J. Andersen, E. Granum, Skin Colour Detection Under Changing Lighting Conditions, 7th Symposium on Intelligent Robotic Systems, Coimbra, Portugal, 1999, 187-195.

[12] M. Soriano, B. Martinkauppi, S. Huovinen, M. Laaksonen, Skin Detection in Video Under Changing Illumination Conditions, IEEE International Conference on Pattern Recognition, Vol.1, Barcelona, Spain, 2000, 839-842.

[13] M. Jones, J. Rehg, Statistical Color Models with Application to Skin Color Detection, Compaq Cambridge Research Lab Technical Report CRL 98/11, 1998.

[14] J.C. Terillon, M.N. Shirazi, H. Fukamachi, S. Akamatsu, Comparative Performance of Different Skin Chrominance Models and Chrominance Spaces for the Automatic Detection of Human Faces in Color Images, IEEE Fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 2000, 54-61.

[15] J. Brand, J.S. Mason, A Comparative Assessment of Three Approaches to Pixel-level Human Skin Detection, IEEE International Conference on Pattern Recognition, Vol.1, Barcelona, Spain, 2000, 1056-1059.

[16] J. Galvin, B. McCane, K. Novins, D. Mason, S. Mills, Recovering Motion Fields: An Evaluation of Eight Optical Flow Algorithms, British machine Vision Conference, Southampton, UK, 1998, 195-204.

[17] J.W. Davis, A.F. Bobick, The Representation and Recognition of Action Using Temporal Templates, IEEE Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 1997, 928-934.

[18] L. Vincent, P. Soille, Watersheds in Digital Spaces: An efficient Algorithm Based on Immersion Simulations, IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(6), 1991, 583-589.

[19] T.F. Cootes, C.J. Taylor, Active Shape Models - 'Smart Snakes', British machine Vision Conference, Leeds, UK, 1992, 266-275.