GEODESIC ACTIVE REGIONS FOR SEGMENTATION AND TRACKING OF HUMAN GESTURES IN SIGN LANGUAGE VIDEOS

Olga Diamanti *and* *Petros Maragos*

School of E.C.E., National Technical University of Athens, Athens 15773, Greece.

olga.diam@gmail.com maragos@cs.ntua.gr

ABSTRACT

Reliable segmentation and motion tracking algorithms are required to achieve gesture detection and tracking for human-machine interaction. In this paper we present an efficient method for detecting and tracking moving hands in sign language video frames. We make use of the geodesic active region framework in conjunction with new color and motion forces; color information is provided by a skin color model, while motion information is derived from the optical flow field. Extensive experimentation indicates that the proposed algorithm behaves sufficiently well for gesture detection and tracking.

Index Terms— Segmentation, Tracking, Geodesic Active Regions, Color Model, Optical Flow, Sign Language

1. INTRODUCTION

Detecting objects (via segmentation) and tracking their motion are major issues in image processing and computer Vision, and are especially important in applications such as Sign Language Recognition(SLR), where the correct detection and understanding of hand motion is perhaps the most significant factor for the interpretation of a particular sign. In this paper we address the problem of signer detection and tracking in SLR applications, and propose an improved method for hand detection and motion tracking. Further, our approach is more general than SLR, and can be used in other image processing applications as well.

Most current approaches for the above tasks either rely on specific equipment (e.g. data gloves), to localize the signer's arms and fingers in space, or on the processing of visual data acquired by cameras. Systems of the first category, while accurately reconstructing the handshapes, are rather unnatural and too expensive to be used in everyday applications. Our algorithm focuses on the analysis of each video frame in order to extract information about the two-dimensional shape of the signer's hands. We use color and motion information to detect the signer and separate the hands. Color is particularly useful when the signer's hands and face are the only skin-colored objects in the image ([1]), while for the motion information to be of advantage, no significant movement of the signer's head, or other objects in the background, should be present ([2]).

Section 2 summarizes the relevant theory about the basic components of our method. In Section 3, the proposed algorithm is presented, while Section 4 contains experimental results.

2. BACKGROUND INFORMATION

In this section we briefly discuss the concepts of segmentation with geodesic active contours and optical flow estimation. These methods will be combined in the next section, in order to track moving hands.

2.1. *Geodesic Active Region Model*

For the segmentation of the video frames we shall use the geodesic active regions (GAR) approach. The GAR model was introduced by Paragios and Deriche in [3], based on the geodesic active contours (GAC) [4] and the region competition [5] approaches. The GAC are deformable two-dimensional contours, which evolve to minimize a suitable energy functional, designed to meet the specific needs of the segmentation process.

More specifically, let C be a planar curve with arclength parameter s and length $L(C)$, and let $\vec{C}(s)$ be its arc-length parametrization. In the GAC model we aim at minimizing the functional $E = \int_0^{L(C)} g(I(\vec{C}(s)))ds$, where I is the intensity image we wish to segment. The function $g : [0, +\infty] \rightarrow \mathbb{R}^2$ is a stopping function, designed to assume minima at image edges, with the property $g(r) \rightarrow 0$ as $r \rightarrow +\infty$. The selected energy functional ensures that the stable state of the curve will satisfy some smoothness criteria and will also tend to locate itself in regions of the image where the image gradient magnitude is relatively large (i.e. on the image edges). The steepest descent method is used for the minimization process, resulting in an Euler-Lagrange PDE for the evolution of the curve.

In order for the numerical solution of this PDE to allow for topological changes to the curve, the GAC model is usually combined with the level-sets method [6]. Thus, the contour C is defined implicitly as the zero level set, at each time point, of an embedding scalar function u defined on the image plane: $C(t) = \{(x, y) : u(x, y, t) = \lambda\}$. A commonly used embedding surface is the signed distance function from the evolving contour. Once we have defined the contour in terms of the embedding function, we can now extend the evolution PDE for the contour to obtain the evolution PDE for the function u:

$$\frac{\partial u}{\partial t} = div\left(g(I)\frac{\nabla u}{\|\nabla u\|} + F(u)\right)\|\nabla u\| \qquad (1)$$

The GAC model can be enhanced with the addition of external forces, represented by $F(u)$ in (1), to the evolution PDE. An example is the GAR model ([3]), in which the image is partitioned into two or more regions, which are assumed to be homogenous with respect to some particular statistically modeled image feature. When the image consists of only two regions, A and A^c, we obtain the following equation for the evolution of u:

$$\frac{\partial u}{\partial t} = div\left[g(I)\frac{\nabla u}{\|\nabla u\|} + \alpha \log\left(\frac{P_A(I)}{P_{A^c}(I)}\right)\right]\|\nabla u\| \qquad (2)$$

This work was supported by the Greek GSRT grant ΕΗΓ-35 DIANOEMA and by the European Union FP6-IST NoE 'MUSCLE' and FET project 'ASPI'.

where $P(A)$ denotes the probability of pixel x belonging to region A, based on the statistical model for this region. In the above equation, which we will extensively use throughout the paper, the evolution is guided by a region based force and an edge based force.

2.2. *Optical Flow Estimation*

A variety of methods have been developed to estimate in the best possible way the optical flow in an image sequence. Well-known among the differential techniques is the Lucas-Kanade estimation method, which we will use in the present paper. In brief, the Lucas-Kanade method is based on the assumption that the gray level of a moving pixel remains unchanged in a short sequence of images, for small pixel displacements and illumination changes: $I(x,y,t) = I(x+\delta x, y+\delta y, t+\delta t)$, which results to the well-known constraint: $\nabla I \cdot \vec{v} = -I_t$, where $v = [V_x, V_y]^T$ is the optical flow vector at pixel (x,y). To solve it, Lucas and Kanade ([7]) proposed to use the least-squares method, by assuming that the optical flow retains the same value inside each pixel's neighborhood. Specifically, by assuming local invariance in a $m \times m$ window centered on each pixel, and assigning an index $i = 1, 2, ..., n = m^2$ to each pixel inside the window, we obtain the least-squares solution $\vec{v} = (A^T A)^{-1} A^T(-b)$, where $A = [\mathbf{I_x}\ \mathbf{I_y}]$, $b = [\mathbf{I_t}]$, and $\mathbf{I_x}, \mathbf{I_y}, \mathbf{I_t}$ are $n \times 2$ column vectors containing the horizontal, vertical and time derivatives of the image intensity at each window pixel.

3. PROPOSED ALGORITHM

3.1. *Skin detection and Segmentation*

Skin-color segmentation is feasible because the human skin has a color distribution that usually differs from that of the background. It is mostly achieved by using only the chrominance (and not the luminance) components of image pixel colors. This ensures that the segmentation algorithm remains, to the best possible extent, unaffected by changes in the scene brightness and by differences in skin color among various users. As a result, the RGB color space is not very efficient for the detection of skin pixels. Instead, popular color spaces used for skin detection include the YCbCr, HSV, CIE Lab spaces ([2],[8]). In our method, the Lab color space is used.

In order to detect skin in an image by using color cues, a skin color model is needed. In our work, we use a relatively simple skin model, since the skin detection phase only serves as the input to the segmentation module, which will be able to segment the skin region with much greater detail. In order to construct this model we manually select and crop training skin regions, from various signers and obtain a large training set. For each cropped region we apply the following two steps: 1) Transform the region pixels from the RGB color space to the Lab color space, and 2) Separate the chrominance values a, b and store the pair of values (a, b) in a suitable two-dimensional accumulator, so that each point in the accumulator corresponds to a specific pair (a, b) and its value to the relative frequency of appearance of this pair in all of the cropped regions. The accumulator array is then smoothed by convolution with a Gaussian function. The resulting array approximates the probability of a certain pair (a, b) belonging to a skin colored region.

3.2. *Combining Skin Color Information with the GAR model*

The intensity image I can be partitioned into two separable regions, one being the union of the skin-colored regions, and the other consisting of the rest of the image pixels, which will be referred to as "background". We may therefore adapt the GAR model to introduce a new force for skin segmentation:

$$F_{color} = \log\left(\frac{P_s(\vec{x})}{P_b(\vec{x})}\right) + cg(I) \tag{3}$$

where P_s, P_b denote the probability of a certain pixel belonging to the skin or background regions, respectively, and $g(I)$ is the edge-detection stopping function of Sec. 2.1.

We estimate the probability P_s via the skin color model we discussed earlier. The background probability is derived straightforwardly as $P_b = 1 - P_s$. Therefore the embedding function evolves according to the following equation:

$$\frac{\partial u}{\partial t} = \left[c_1 div\left(g(I)\frac{\nabla u}{\|\nabla u\|}\right) + c_2 \log\left(\frac{P_s(\vec{x})}{1 - P_s(\vec{x})}\right) + c_3 g(I)\right]\|\nabla u\| \tag{4}$$

in which the reader may recognize the typical edge-based force, our proposed color force and a third force, known as "balloon force" ([4]) which speeds up the evolution procedure by attempting to minimize the surface of the embedded curve.

Our proposed force ensures that the curve will eventually converge to those image edges that separate skin regions from the background. The use of the GAR framework eliminates any issues concerning the continuity of the skin regions detected by the color model, as the smooth curve will enclose the whole skin region, provided that these discontinuities are not too large. On the other hand, the statistical force makes the segmentation model more robust with respect to weak or false intensity edges.

3.3. *Incorporation of the Optical Flow Force*

While color information is a major cue for the detection of hands in images, motion information is equally crucial when the goal is to recognize human gestures. As far as SLR is concerned, it may help to resolve the well-known problems arising in the presence of hand-face or hand-hand occlusions, given that the signer's face remains relatively motionless while the hands move vividly. In this paper we introduce a way to exploit the available motion information by using, once again, the GAR model. Namely, we will again use a statistical force of the logarithmic form presented in section 2.1, with the image now being partitioned into a moving and a static region.

The motion information is provided from the optical flow field, derived with the aforementioned Lucas-Kanade algorithm. The magnitude of the optical flow field could be used to obtain the probability of a certain image pixel belonging to either the static or the moving component. Thus, we introduce a new evolution force for the active contour, according to the following steps:

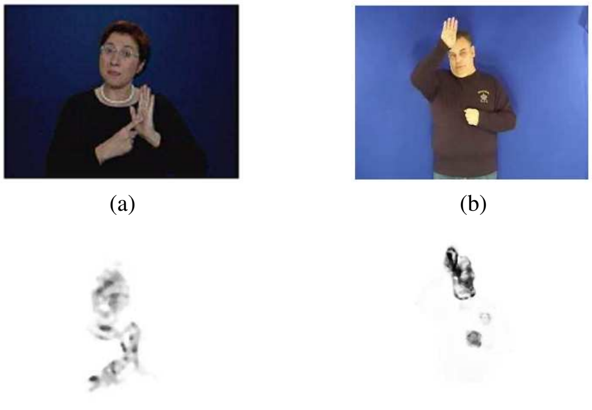
1. Estimation of the optical flow field OF and its magnitude $|OF(x,y)| = \sqrt{V_x^2 + V_y^2}$ and of the moving region probability $P_{mov}(x,y) = |OF(x,y)|/\max(|OF|)$
2. Motion force: $F_{mov}(x,y) = \log\left(\frac{P_{mov}(x,y)}{P_{stat}(x,y)}\right)$

where $P_{stat}(x,y) = 1 - P_{mov}(x,y)$ is the probability of the pixel at location (x,y) belonging to the static region ("background").

This new motion force operates in an analogous way to the color force. In leads the evolving contour to converge so as to include the locations where motion is detected. This could be used to locate the hands in an image and discriminate them from the face region, which will also be detected by the skin color model. In such a way, we can diminish or eliminate errors in the estimation of the hands' positions in the presence of occlusions. The overall force at point $\vec{x}$ is:

$$F(\vec{x}) = \log\left(\frac{P_{skin}(\vec{x})}{P_{nonskin}(\vec{x})}\right) + \log\left(\frac{P_{mov}(\vec{x})}{P_{stat}(\vec{x})}\right) + cg(I(\vec{x})) \quad (5)$$

A similar method to utilize motion information was proposed in [9]. However, the algorithm presented in this paper differs in two major points: firstly, the authors of [9] use the difference between two consecutive images to estimate motion, by statistically modeling the difference image. In this paper we use the optical flow field, which provides a much more accurate and robust way to detect motion. Also, the model in [9] is edge-based; it relies on the computation of an image containing the edges between moving and static image regions. The algorithm analyzed here is region-based and requires only an estimation of the moving and static probabilities.


(a) (b)

(c) (d)

Fig. 1. Skin Detection for 2 different signers. Top row: Initial images, Bottom row: Results of the skin detection module.

4. EXPERIMENTAL RESULTS

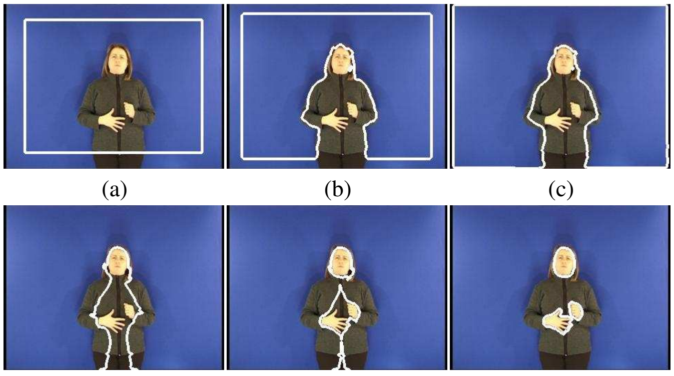
The algorithms presented were tested with a large set of video frames, for several different signs and four different signers, to evaluate their performance in various illumination and skin color conditions. The videos on which our algorithm was tested were derived from a fully annotated Greek Sign Language Corpus ([10]), created to support the purposes of human-computer interaction.

4.1. *Skin detection*

Figure 1 depicts the results of applying the skin color model to two different input images. The model succeeds with relative accuracy to detect skin regions. Some vagueness in the detection of the regions is inevitable, and some of the regions may appear to be scattered and discontinuous. However, no further processing is needed, as these imperfections will be resolved in the final segmentation stage.

4.2. *Image Segmentation using Color Information*

In Fig. 2 we present four different stages of the curve's evolution, resulting from the application of the GAR model with the color force to the intensity images. The results for some input images are shown in Fig. 3. The evolving curve was initialized near the borders of the image, thus enclosing all skin colored regions. The GAR model was numerically implemented using the fast multigrid scheme of [11], achieving a segmentation time of 0.2-0.3 sec/frame. It should be noted that the weights c_1, c_2, c_3 in eq. (2) were kept the same for all experiments, which demonstrates that the algorithm is to a large extent signer- and illumination-independent.


(a) (b) (c)

(d) (e) (f)

Fig. 2. The evolution of the geodesic curve.(a):initial position of the contour, (b)-(e):intermediate positions, (f):final position.

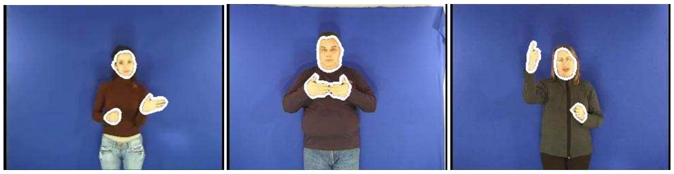

(a) (b) (c)

Fig. 3. Segmentation with GARs and color force. The final position of the evolving curve is dilated for visualization purposes.

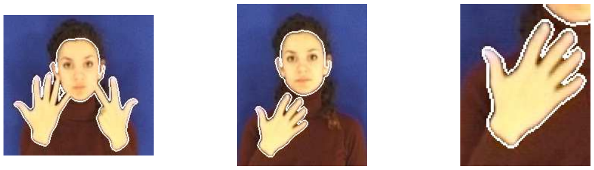


Fig. 4. Details of the segmentation results.

In order to demonstrate the ability of the proposed method to successfully segment the signers' hands and finger, we provide in Fig. 4 two close-up views of our segmentation results. Also, in Fig. 5 we compare our method to three other well-known segmentation approaches. The proposed method succeeds in detecting the desired objects only (notice the false regions detected in Fig. 5(b) and Fig. 5(c)) and giving a smooth contour, while requiring significantly less training for the color model (compared to Fig. 5(a)).

4.3. *Hand Detection and Tracking using the Optical Flow Field*

In Fig. 6 we present results of the application of our algorithm in order to detect moving hands in sign language video frames. The algorithm succeeds in the detection of the moving hand, even in presence of occlusions (as in Fig. 6(d)-(f)).

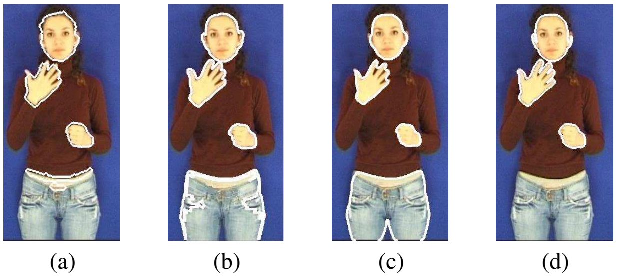


Fig. 5. Comparison of 4 different segmentation methods. (a) Skin Color Segmentation without GACs ([1]), (b) GACs with edge force only ([4]), (c) Chan-Vese ([12]), (d) GARs with color force.

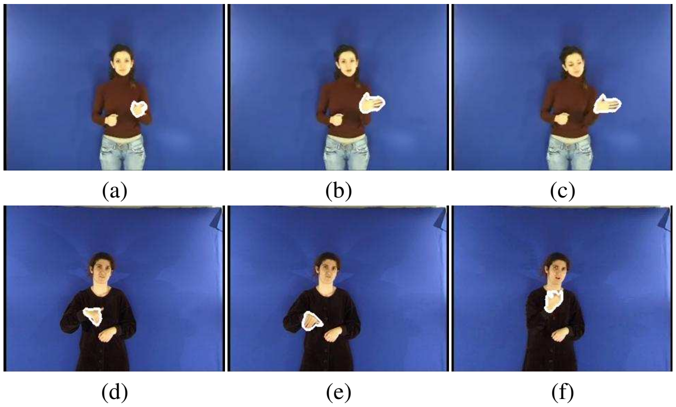


Fig. 6. Moving Hand Detection using both color and motion cues. Signs : "to the left" and "to leave".

4.4. *Further Experimentation*

As part of our efforts to cope with the issue of occlusions in SL analysis, we also experimented with a different approach, which makes use of prior information ([13],[14],[15]) about the hand's shape to locate the hand in a sequence of frames. We make the assumption that a hand's shape does not significantly change between consecutive frames, and can therefore be used as a prior. The shape is represented by its signed distance transform, which undergoes geometric transformations together with the evolving curve until the curve converges to the position where the shape is best aligned. Some early but very promising results are shown in Fig. 7. The same handshape is used as a prior for frames extracted from different signers.

5. CONCLUSIONS

We have developed an efficient algorithm for the segmentation of moving hands in image sequences, based on the GAR framework, with a new way of incorporating color and motion information. This algorithm constitutes a promising approach towards hand detection and tracking for SLR applications. Motion and color are treated simultaneously and in a unified way, so that the curve evolution can take place in a single step. The proposed method could be incorporated into the visual front-end of a video analysis system for SLR. Future work includes extending our method to cope with uncontrolled cluttered environments, and reducing the processing time.

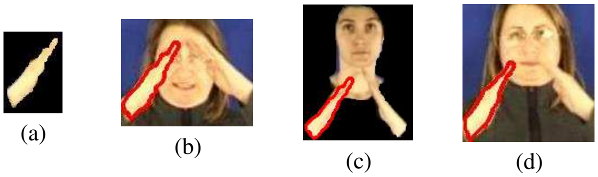


Fig. 7. Hand segmentation with shape prior. (a): prior, (b)-(d): results for skin regions extracted from the signs "tent" and "house".

6. REFERENCES

[1] A.A. Argyros and M.I.A. Lourakis, "Real-Time Tracking of Multiple Skin-Colored Objects with a Possibly Moving Camera," in *Proc. Eur. Conf. Comp. Vision*, 2004, vol. III.

[2] N. Soontranon, S. Aramvith, and T.H.Chalidabhongse, "Improved Face and Hand Tracking for Sign Language Recognition," *Proc. Int'l Conf. on Information Technology*, 2005.

[3] N. Paragios and R. Deriche, "Geodesic Active Regions: A New Framework to Deal with Frame Partition Problems in Computer Vision," *Journ. of Vis. Commun. and Image Repres.*, vol. 13, no. 1/2, pp. 249–268, 2002.

[4] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic Active Contours," *Int'l J. Computer Vision*, vol. 22, no. 1, pp. 61–79, 1997.

[5] S.C. Zhu and A. Yuille, "Region Competition: Unifying Snakes, Region Growing and Bayes/MDL for Multiband Image Segmentation," *TPAMI*, vol. 18, no. 9, pp. 884–900, 1996.

[6] S. Osher and J.A. Sethian, "Fronts Propagating with Curvature Dependent Speed:Algorithms Based on Hamilton-Jacobi Formulations," *Journ. of Comp. Phys.*, vol. 79, pp. 12–49, 1988.

[7] B.D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in *Proc. of Imaging understanding workshop*, 1981, pp. 121–130.

[8] H. Hongo, M. Ohya, M. Yasumoto, and K. Yamamoto, "Face and Hand Gesture Recognition for Human-computer Interaction," in *ICPR*, 2000, pp. Vol II: 921–924.

[9] R. Deriche and N. Paragios, "Geodesic Active Contours and Level Sets for the Detection and Tracking of Moving Objects," *IEEE TPAMI*, vol. 22, no. 4, pp. 415, April 2000.

[10] E. Efthimiou and E. Fotinea, "GSLC: Creation and annotation of a Greek Sign Language Corpus for HCI," in *Proc. 4th Int'l Conf. on Univ. Access in HCI*, 2007.

[11] G. Papandreou and P. Maragos, "Multigrid Geometric Active Contour Models," *IEEE Trans. on Image Processing*, vol. 16, no. 1, pp. 229–240, January 2007.

[12] T. Chan and L. Vese, "Active contours without edges," *IEEE Trans. on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.

[13] D. Cremers, F. Tischhauser, J. Weickert, and C. Schnorr, "Diffusion Snakes: Introducing Statistical Shape Knowledge into the Mumford-Shah Functional," *Int'l J. Computer Vision*, vol. 50, no. 3, pp. 295–313, December 2002.

[14] M.E. Leventon, W.E.L. Grimson, and O.D. Faugeras, "Statistical Shape Influence in Geodesic Active Contours," in *Proc. Int'l Conf. on Comp. Vis. & Patt. Rec.*, 2000, pp. I: 316–323.

[15] M. Rousson and N. Paragios, "Shape Priors for Level Set Representations," in *Proc. Eur. Conf. Comp. Vision*, 2002.

[16] Ch.G. Bergeles, "Tracking Moving Objects," *Diploma Thesis*, NTUA 2006.