KERNEL-BASED HEAD TRACKER FOR VIDEOPHONY

Ilse Ravyse, Valentin Enescu and Hichem Sahli

Vrije Universiteit Brussel, ETRO Department
Interdisciplinary Institute for Broadband Technology
Pleinlaan 2, 1050 Brussel
{icravyse, venescu, hsahli}@etro.vub.ac.be

ABSTRACT

An approach for automatically segmenting and tracking a face in a sequence of color images is presented. The initial face localization consists of a two-step process: the face candidates selection using skin color clustering, and the face verification which picks the best face candidate based on shape and color cues. The tracking of the detected head in the subsequent image frames is performed via a kernel-based method wherein a joint spatial-color probability density characterizes the head region. The parameterized motion and the illumination changes affecting the target are estimated by minimizing the l_2 distance between the densities of the head candidate and the head model. The proposed algorithms achieve reliable detection and tracking results.

1. INTRODUCTION

Head detection and tracking schemes have received much attention from the image processing community due to their wide range of applications. Indeed, they contribute to a substantial amelioration of multimedia applications such as videophony, tele-presence, character animation, user-friendly interfaces and access-security via person identification [1]. The task of finding a person's face in an image is referred to as face localization, face extraction or face segmentation. Several approaches have been proposed. The grouping of facial features into face candidates, the use of heuristic rules about a typical face and the correlation with a fixed or statistical face template [2].

Recently, kernel-based tracking in the spatial-feature domain has emerged as a robust and accurate method [3, 4]. This tracking approach affords us to circumvent the high computational burden of the face segmentation techniques. Additionally, since illumination variations often occur in videophony, we explicitly model them.

The paper is organized as follows. Section 2 presents the initial face segmentation. In section 3 the head tracking algorithm is elaborated and experimental results are reported. Finally some conclusions are given in section 4.

2. HEAD DETECTION

2.1. Face Candidates Selection

The skin color is a powerful descriptor for extracting the human face. Indeed, the human face has a specific color distribution that often differs from that of the background objects.

Following Chai *et al.* [5], we employ the YC_rC_b color space to describe the skin color distribution. Thus, a skin color map is created by assigning to each color of the skin pixels the number of occurrences in a face database [6]. The skin-color region in the C_rC_b space is then confined to a polygon that is adapted to each luminance level Y [7]. Figure 1f shows the polygonal skin domain on a C_rC_b plane.

Having defined the skin color domain enables us to label the image pixels as skin/non-skin. To avoid problems with noise and misclassifications, we first perform a watershed segmentation [8, 9, 6], and then classify each segment. The latter is achieved by detecting whether the segment's mean color resides within the pre-determined skin color domain. The face candidate regions are then obtained by successively merging the neighboring skin segments. However, the existence of background regions which have a skin-like color can be misleading. Indeed, face candidates including the actual face and some background regions can arise. It occurs frequently for persons with a skin-colored hair. To alleviate this problem, a supervised clustering procedure is applied before merging. The approach is similar to that discussed by Chai *et al.* [5], but we incorporate direct face knowledge by assuming that the face skin is more red than other body parts [10]. The clustering is performed via the seeded k-means algorithm, applied on the C_rC_b features of the skin segments. Three seeds are used for initialization: a first one at the reddish top of the polygon (K_1), a second one at the bluish end of the polygon (K_2), and a third one (K_3) at the crossing of the bounding lines of $R/G > 1$ and $R/B > 1$ semiplanes. K_3 has been added to deal with outliers and yellow/black skin. Finally, the face candidate regions are obtained by successively merging the neighbor-

ing segments belonging to K_1, K_2, and K_3.

Figure 1 shows the clustering of the skin color map for the *Claire* image and the face candidate regions.

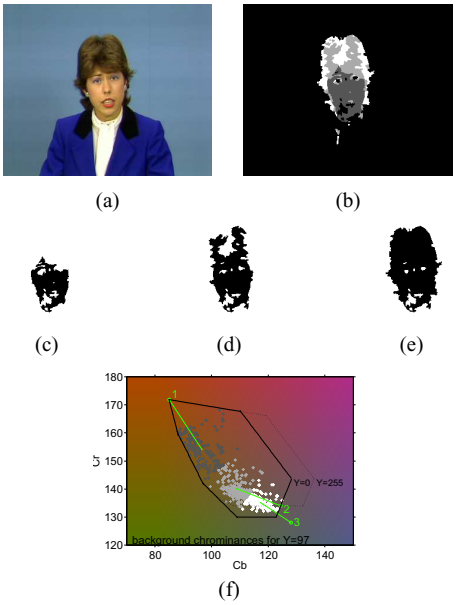


Fig. 1. *Claire*: (a) face image; (b) labeled image; (c, d, e) face candidates $\{R^i\}_{i=1}^3$; (f) k-means clustering of the skin color map with seed and final mean numbered by cluster

2.2. Face Verification

Having detected several face candidate regions $\{R^i\}_{i=1}^N$, the face verification step picks the best candidate as the face region. This is done by quantifying two shape cues (Q_1 and Q_2) based on the best-fit ellipse E^i [11] bounding each candidate. E^i is described by its center, orientation, and the lengths of major and minor axes: $E^i = (x^i, y^i, \theta^i, a^i, b^i)$. Additionally, two quantified facial feature cues (Q_3 and Q_4) based on the gray-value Y are specified. These cues are defined as follows:

- Q_1 checks if the height to width ratio of the ellipse approximates that of a human face:
  $Q_1(R^i) = 1/|(a^i/b^i) - 1.5|$
- Q_2 is large if the ellipse has a high fill percentage:
  $Q_2(R^i) = \text{card}(R^i \bigcap E^i)/\text{card}(E^i)$
- Q_3 expresses the gray-value variance:
  $Q_3(R^i) = 1/\text{var}^Y_{R_i \bigcup E_i}$
- Q_4 allows only a low number of strong corners. The corner image C is obtained by applying an isodata threshold on the outcome of a pixel-wise corner measure acting on the gray-value image [12]:
  $Q_4(R^i) = 1/\text{card}(R^i \bigcap C)$

The quantified cues Q_j of each face candidate R^i can be compared via a modified z-score:

$$z_j^i = \frac{Q_j(R^i)}{\sqrt{\left(\sum_{k=1}^N Q_j(R^k)\right)/N}} \quad (1)$$

The minimal z-score for each cue is determined. Finally, the face measure M^i for a face candidate i is formulated as

$$M^i = \sqrt{\sum_{j=1}^4 \left(z_j^i - \min_{k=1,\ldots,N}(z_j^k)\right)^2} \quad (2)$$

The face candidate that has the maximal measure localizes the face in the image. Figure 2 shows the selected face candidate for the *Claire* image.

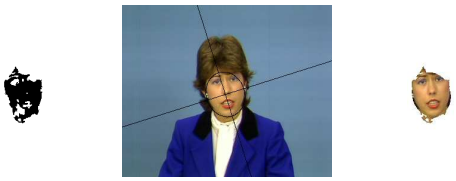


Fig. 2. *Claire*: (a) selected face candidate R^i; (b) ellipse E^i; (c) detected face $R^i \bigcup E^i$

3. HEAD TRACKING

The head (target) model consists of the coordinates $\mathbf{x}_i = (x_i, y_i)$ and the colors $\mathbf{u}_i = (R, G, B)$ of the N pixels inside the detected ellipse head shape in the first image frame (see section 2). In the next frames, the target is subject to motion and illumination changes, which are described by parametric models. Head tracking is then cast as an optimization problem where the parameters are estimated using the similarity between the hypothesized and the candidate target. By hypothesized target we mean the samples $\{\mathbf{y}_i, \mathbf{v}_i\}_{i=1}^N$ obtained by transforming $\{\mathbf{x}_i, \mathbf{u}_i\}_{i=1}^N$ via the parametric models, while the candidate target consists of the samples $\{\mathbf{y}_j = (x_j, y_j), \mathbf{v}_j\}_{j=1}^M$ inside the elliptic image-region resulting from the motion of the initial ellipse head shape. Further, based on these samples, spatial-color kernel-based probability density estimates are build and their similarity is assessed via an l_2 norm. We next describe the two parametric models and then derive the formulas for parameter estimation.

3.1. Parametric Motion Model

At each time step or image frame, the target undergoes a geometric transformation. We assume it consists of a rotation by an angle θ around a fixed center point $\mathbf{x}^*$ and translation to a new center point $\mathbf{y}^*$:

$$\mathbf{y} = \mathbf{M}_\theta\,(\mathbf{x} - \mathbf{x}^*) + \mathbf{y}^* \quad (3)$$

with $\mathbf{M}_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ being the rotation matrix.

3.2. Illumination Model

The illumination of the head can change e.g. when it is oriented differently with respect to the light source. The *Bidirectional Reflectance Distribution Function* ($BRDF$) is a scalar mathematical function specifying the percentage of light arriving from each incoming direction (angle to light source) that is reflected in an outgoing direction (viewing angle to the camera) [13]. We can assume that both directions are constant, so the illumination level is modified by a contrast constant λ. To take into account reflections from neighboring object and camera automatic gain, a brightness constant δ has to be added [14]. The illumination model applied on the target color model gives the hypothesized target color

$$\mathbf{v} \quad = \quad \lambda \mathbf{u} + \delta \tag{4}$$

3.3. Kernel-based Distance Optimization

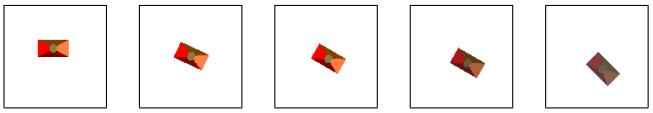
Optimizing the correlation between the hypothesized and the candidate target is done by minimizing the distance D between their representations with respect to the parameter vector $\psi = (\mathbf{y}^*, \theta, \lambda, \delta)$:

$$\hat{\psi} = \arg\min_{\psi} D((\mathbf{x}_i, \mathbf{u}_i)_{i=1}^N, I(t), \psi) \tag{5}$$

In our case, the distance is the l_2 norm distance between the joint spatial-color probability densities [3], as position and color are considered random variables. The kernel density estimation is a non-parametric technique to estimate the probability density distribution from the samples of a multivariate random variable [4]. The kernel density estimate p of a target with N samples $\{\mathbf{x}_i, \mathbf{u}_i\}_{i=1}^N$ is

$$\begin{aligned} p(\mathbf{x}, \mathbf{u}) &= p(x, y, u^{(1)} = R, u^{(2)} = G, u^{(3)} = B) \\ &= \frac{\alpha}{N} \sum_{i=1}^{N} \left(K_{h_x}(\mathbf{x} - \mathbf{x}_i) \prod_{l=1}^{3} K_{h_u}(u^{(l)} - u_i^{(l)}) \right) \end{aligned} \tag{6}$$

where $K_h(\mathbf{y}) = \frac{1}{(\sqrt{2\pi}h)^d} \exp\left(-\frac{1}{2}\left(\frac{\|\mathbf{y}\|^2}{h^2}\right)\right)$ is a gaussian kernel and α is a normalization constant. The spatial kernel has a bandwidth h_x and dimension $d_x = 2$ and the color kernels have a bandwidth h_u and $d_u = 1$. The candidate target density $p(\mathbf{y}, \mathbf{v})$ uses (6) on the samples $\{\mathbf{y}_j, \mathbf{v}_j\}_{j=1}^M$ inside the ellipse with new center $\mathbf{y}^*$ and rotated by θ around $\mathbf{x}^*$ of (3). The target model samples are transformed by (3) and (4) to obtain the hypothesized set of samples in the current frame, $\{\mathbf{y}_i, \mathbf{v}_i\}_{i=1}^N$, from which we estimate via (6) the hypothesized target density, $p(\mathbf{y}, \mathbf{v}; \psi)$. The l_2 distance (5)



(1) (2) (3) (4) (5)

Fig. 3. The *Synthetic* sequence frames

between the hypothesized and the candidate target densities is given by

$$D = \int \|p(\mathbf{y}, \mathbf{v}; \psi) - p(\mathbf{y}, \mathbf{v})\|^2 d\mathbf{y} d\mathbf{v} = a \sum_{i,j=1}^{N,M} w_{ij} + b$$

where $w_{ij} \triangleq K\left(\frac{\mathbf{M}_\theta\ (\mathbf{x}_i - \mathbf{x}^*) + \mathbf{y}^* - \mathbf{y}_j}{\sqrt{2}}\right) \prod_{l=1}^{3} K\left(\frac{\lambda u_i^{(l)} + \delta - v_j^{(l)}}{\sqrt{2}}\right)$ and a and b are constant. Minimizing D is accomplished by setting to zero its componentwise derivatives with respect to ψ. The estimated parameters are obtained via an iterative update scheme $\psi^{(n+1)} = \mathbf{F}(\psi^{(n)})$, where $\mathbf{F}$ is given by (7)-(10):

$$\mathbf{y}^{*(n+1)} \quad = \quad \frac{\sum_{i,j} w_{ij}^{(n)} (\mathbf{y}_j - \mathbf{M}_{\theta^{(n)}}\ \tilde{\mathbf{x}}_i)}{\sum_{i,j} w_{ij}^{(n)}} \tag{7}$$

$$\theta^{(n+1)} \quad = \quad \arctan\left(-\frac{\sum_{i,j} w_{ij}^{(n)} (\tilde{y}_i \bar{x}_j - \tilde{x}_i \bar{y}_j)}{\sum_{i,j} w_{ij}^{(n)} (\tilde{x}_i \bar{x}_j + \tilde{y}_i \bar{y}_j)} \right)$$

$$\text{with } \bar{\mathbf{y}} = \mathbf{y} - \mathbf{y}^*;\ \tilde{\mathbf{x}} = \mathbf{x} - \mathbf{x}^* \tag{8}$$

$$\lambda^{(n+1)} \quad = \quad \frac{\sum_{i,j} w_{ij}^{(n)} \sum_{l=1}^{3} (v_j^{(l)} - \delta^{(n)}) u_i^{(l)}}{\sum_{i,j} w_{ij}^{(n)} \sum_{l=1}^{3} (u_i^{(l)} u_i^{(l)})} \tag{9}$$

$$\delta^{(n+1)} \quad = \quad \frac{\sum_{i,j} w_{ij}^{(n)} \sum_{l=1}^{3} (v_j^{(l)} - \lambda^{(n)} u_i^{(l)})}{3 \sum_{i,j} w_{ij}^{(n)}} \tag{10}$$

Note that $\psi^{(0)}$ is initialized with the parameters estimated at the previous frame. Iterating stops when the change in the distance D is under a preset threshold, thereby yielding the final parameters $\hat{\psi}$.

3.4. Results

Synthetic Sequence: The algorithm (7)-(10) is applied with $h_x = 3$ on the sequence shown in Figure 3 (illumination changes are present in frames 4 and 5). Figure 4 presents the ground truth vs. the tracking results with and without modeling the illumination. The final target estimates in the last frame are depicted in Figure 5. When not using the illumination model, we try to cope with illumination changes by increasing the bandwidth of the color kernel. Note that, this leads to poor estimates as compared with explicitly modeling the illumination changes.

Videophony Sequence: The *Claire* sequence is processed starting from the detected ellipse comprising the face (see

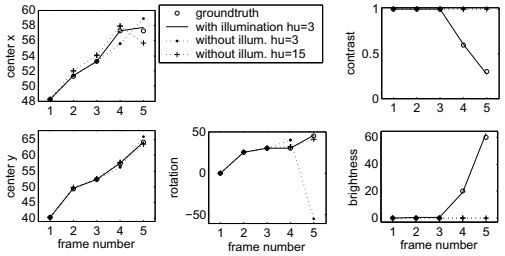
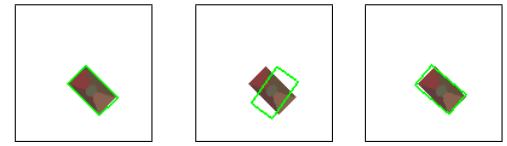


Fig. 4. Tracking results for the *synthetic* sequence



(with illumination $h_u = 3$) (without illum. $h_u = 3$) (without illum. $h_u = 15$)

Fig. 5. Final target estimate for the *synthetic* sequence - frame 5

Figure 2b). Tracking proceeds with $h_x = h_u = 3$. The parameter estimates ($\hat{\mathbf{y}}^*$ and $\hat{\theta}$) are presented in Figure 6. The final head estimate of frame 65 is shown in Figure 7. The tracking recovers well from the 3-dimensional motion of the face (when *Claire* is looking aside or down) around frame 25 and 55.

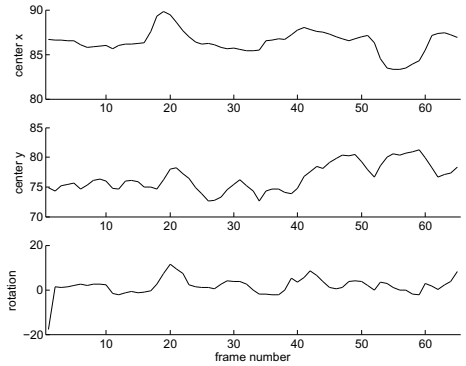


Fig. 6. Tracking results for the *Claire* sequence

4. CONCLUSION

In this paper, we presented a head detection and tracking framework for videophony. Head detection is carried out by labeling the skin-colored segmented image regions via clustering the pixel values in the YC_rC_b color space. The use of a polygonal boundary for skin colors avoids the storage of a human face database or skin color probability histograms as in [6]. Combining several shape and facial feature cues ensures an adequate detection of the face. This automated head detection is used to initialize the head tracker for which the kernel-based approach proved to be robust to the 3-dimensional motion of the face. Moreover, incorporating an illumination model into the tracking equations

Fig. 7. Final head estimate for the *Claire* sequence - frame 65; used target model in upper left corner

enables us to cope with potentially distracting illumination changes.

Acknowledgement: This research has been conducted within the framework of the Inter-Universitary Attraction-Poles program number IAP 5/06 Advanced Mechatronic Systems, funded by the Belgian Federal Office for Scientific, Technical and Cultural Affairs.

5. REFERENCES

[1] F. I. Parke and K. Waters, *Computer Facial Animation*. A K Peters, 1996, ISBN 1-56881-014-8.

[2] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, Jan. 2002.

[3] H. Zhang, Z. Huang, W. Huang, and L. Li, "Kernel-based method for tracking objects with rotation and translation," in *17th International Conference on Pattern Recognition (ICPR'04), Cambridge UK, August 23 - 26, 2004*, vol. 2, Aug. 2004, pp. 728– 731.

[4] D. Comaniciu, "Bayesian kernel tracking," in *Annual Conf. of the German Society for Pattern Recognition (DAGM'02), Zurich, Switzerland, September 16-18 2002*, Sep. 2002, pp. 438–445.

[5] D. Chai and K. Ngan, "Face segmentation using skin-color map in videophone applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 551–564, Jun. 1999.

[6] B. Martinkauppi, M. Soriano, and M. Pietikainen, "Detection of skin color under changing illumination: a comparative study," in *12th International Conference on Image Analysis and Processing (ICIAP 2003), Mantova, Italy*, Sep. 2003, pp. 652–657.

[7] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Transactions on Multimedia*, vol. 1, no. 3, pp. 264–277, Sep. 1999.

[8] I. Pratikakis, H. Sahli, and J. Cornelis, "Low level image partitioning guided by the gradient watershed hierarchy," *Signal Processing*, vol. 75, pp. 173–195, 1999.

[9] B. Menser and M. Wien, "Segmentation and tracking of facial regions in color image sequences," in *SPIE Visual Communications and Image Processing, Perth, Australia*, vol. 4067, Jun. 2000, pp. 731–740.

[10] J. Brand and J. S. Mason, "A comparative assessment of three approaches to pixel-level human skin-detection," in *15th International Conference on Pattern Recognition, ICPR2000 (September 3-8, 2000) Barcelona, Spain*, vol. 1, Sep. 2000, pp. 1056–1059.

[11] K. Sobottka and I. Pitas, "A novel method for automatic face segmentation, facial feature extraction and tracking," *Signal Processing: Image Communication*, vol. 12, no. 3, pp. 263–281, Jun. 1998.

[12] C. Achard, E. Bigorgne, and J. Devars, "A sub-pixel and multispectral corner detector," in *15th International Conference on Pattern Recognition, ICPR2000 (September 3-8, 2000) Barcelona, Spain*, vol. 3, Sep. 2000, pp. 971–974.

[13] B. Draper and J. R. Beveridge, "Bidirectional reflectance distribution function: Phong reflectance," in *CVonline: On-Line Compendium of Computer Vision [Online]*. R. Fisher (ed)., 2002.

[14] H. Jin, P. Favaro, and S. Soatto, "Real-time feature tracking and outlier rejection with changes in illumination," *IEEE Proceedings Conference on Computer Vision (ICCV), Vancouver, B.C., Canada , July 07 - 14, 2001*, vol. 1, pp. 684–689, Jul. 2001.