Face Tracking for H.264 Encoded Video Sequences

Bogdan Kwolek
Rzeszów University of Technology
W. Pola 2, 35-959 Rzeszów, Poland
bkwolek@prz.rzeszow.pl

Abstract— **In this paper we present a face detection and tracking-based scheme to enhance the perceptual quality of face regions in face-to-face teleconference applications using H.264. The tracking is realized using a particle filter. The weights of the particles are determined on the basis of intensity gradient near the edge of the ellipse, appearance difference and Bhattacharyya distance between weighted histograms. We assign smaller weights to particles which are farther away from the centers of detected faces. We employ a reliability factors of the utilized cues to control the noise variance as well as to supervise the number of particles. After classifying the macroblocks into face and non-face categories we perform a quantization. This requires some modifications in H.264 encoder but requires no modifications to the decoder.**

I. Introduction

The H.264/AVC video coding standard has been developed to accomplish considerable improvements over the existing standards in the compression performance. It provides an increase in compression efficiency of up to 50% over a broad range of transmission rates and image sizes. The video coding still uses the concept of block-based motion compensated prediction to remove temporal correlations and transform-based residual coding to remove spatial redundancies [15]. Each frame of an input video stream is divided into macroblocks. H.264 encodes the video without specific knowledge of the semantic content of the frames and assigns equal importance to each image block. It particular, H.264-based video coding does not pay any special attention to facial regions in order to obtain better perceptual quality at the sacrifice of the quality of non-face regions.

The overall perceptual quality of video transmitted at low bitrates can be improved by encoding the face region with higher bitrate that less relevant background regions [2][6][12]. In order to provide by decoder the perceptually pleasing images where faces are sharper than the background a skin-based face segmentation method has been applied in work [2]. The work [8] also proposed a skin-color based face detection approach and utilized a dynamic weighting adjustment scheme for preferential enhancing the face regions by dropping the static non-face regions. However, skin-like background can lead to many false detections and in consequence to reduction of the efficiency of the coder. An ellipse fitting technique has been used to extract face regions in a teleconferencing application [6]. Approaches that rely solely on intensity information and elliptical shape frequently fail because elliptical shape features appear in multiple non-facial structures. A fast algorithm to detect face regions directly in MPEG video streams has been described in [14]. To enhance the visual quality of the face regions a visual sensivity-based quantization scheme has been proposed in work [4]. This method is incompatible with H.263 standard because the description of the region of interest needs to be extra transmitted to the decoder.

In this paper we present an approach which incorporates the prior knowledge about faces into H.264 encoder to improve the quality of video through selective quantization. We use face detection and particle-based tracking algorithms to locate a rectangular box containing face. After classifying the macroblocks into face and non-face regions we perform a selective quantization. The quantization step is where a significant compression takes place. In H.264-based encoder a fifty-two different quantization step sizes can be chosen and applied to each block separately. The quantization step size is chosen by so called quantization parameter which supports 52 different quantization coefficients. An increment of QP by 1 results in an increase of the required data rate of approximately 12.5% [1][15].

Fast and robust face tracking in an image sequence is highly desirable capability for many multimedia interfaces. Face tracking permits background regions of the image to be discarded and allows the algorithm to concentrate on desirable object-like regions. The tracking provides a focus-of-attention mechanism [12]. To reliably track a face in video sequences we fuse color and shape within a particle filter-based framework. The tracked face is represented by a weighted histogram carrying information about the color and the shape. The histograms are compared using the Bhattacharyya distance. The color distribution is extracted in interior of the ellipse modeling the outline of the tracked head. The color histogram and the parameters of the ellipse are updated over time. To realize robust visual tracking and recognition we incorporate into particle filter an appearance model of the face. Relying on face detection results we assign smaller weights to particles which are farther away from the centers of detected faces. We employ the reliability

0-7803-9134-9/05/$20.00 ©2005 IEEE

factors of the utilized cues to set the noise variance as well as to control the number of particles. Thanks to face detection the system automatically initializes without user intervention, and can reinitialize when the tracking is lost.

The paper is organized as follows. In the next section we briefly outline particle filtering. Section 3 presents the implementation of the particle filter. The quantization strategy is explained in section 4. Section 5 illustrates the performance of the tracking algorithm. The paper is ended with some concluding remarks.

II. PARTICLE FILTERING

The effectiveness of object tracking in image sequences has been greatly improved with the development of particle filtering. The particle filter is an algorithm for estimating the posterior state of a dynamic system over time where the state cannot be measured directly, but may be estimated at the current time-step t. Particle filters are attractive for nonlinear models, multi-modal, non-Gaussian or any combination of these models for several reasons. They utilize imperfect observation and motion models and incorporate noisy collection of observations through Bayes rule. The ability to represent multimodal posterior densities allows them to globally localize as well as relocalize the object of interest in case of failure during tracking. Particle filters are any-time because by supervising the number of samples online they can adapt to the available computational resources.

Two important components of each particle filter are motion model $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ describing the state propagation and observation model $p(\mathbf{z}_t \mid \mathbf{x}_t)$ describing the likelihood that a state $\mathbf{x}_t$ causes the observation $\mathbf{z}_t$. Starting with a weighted particle set $S = \left\{(\mathbf{x}_{t-1}^{(n)}, \pi_{t-1}^{(n)}) \mid n = 1...N\right\}$ approximately distributed according to $p(\mathbf{x}_{t-1} \mid \mathbf{z}_{1:t-1})$ the filter operates through predicting new particles from a proposal distribution. To give a new particle representation $S = \left\{(\mathbf{x}_t^{(n)}, \pi_t^{(n)}) \mid n = 1...N\right\}$ of the posterior density $p(\mathbf{x}_t \mid \mathbf{z}_{1:t})$ the weights of particles are set to $\pi_t^{(n)} \propto \pi_{t-1}^{(n)} p(\mathbf{z}_t \mid \mathbf{x}_t^{(n)}) p(\mathbf{x}_t^{(n)} \mid \mathbf{x}_{t-1}^{(n)}) / q(\mathbf{x}_t^{(n)} \mid \mathbf{x}_{t-1}^{(n)}, \mathbf{z}_t)$. When the proposal distribution from which particles are drawn is chosen as the distribution conditional on the particle state at the previous time step, the importance function reduces to $q(\mathbf{x}_t^{(n)} \mid \mathbf{x}_{t-1}^{(n)}, \mathbf{z}_t) = p(\mathbf{x}_t^{(n)} \mid \mathbf{x}_{t-1}^{(n)})$ and the weighting function takes the form $\pi_t^{(n)} \propto p(\mathbf{z}_t \mid \mathbf{x}_t^{(n)})$. This simplification leads to a variant of a particle filter, CONDENSATION [7]. From time to time the particles should be resampled according to their weights to avoid degeneracy [5].

III. STATE SPACE AND OBSERVATION MODEL

The observation model integrates three different visual cues. In this section we present the motion model and demonstrate how we construct the adaptive observation model.

A. State space and dynamics

The outline of the head is modeled in the 2D-image domain as a vertical ellipse that is allowed to translate and scale subject to a dynamical model. The object state is given by $\{x, \dot{x}, y, \dot{y}, s_y, \dot{s}_y\}$, where $\{x, y\}$ denotes the location of the ellipse center in the image, $\dot{x}$ and $\dot{y}$ are the velocities of the center, s_y is the length of the minor axis of the ellipse and $\dot{s}_y$ is the rate at which s_y varies. We use a first-order auto-regressive dynamic model $\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{w}_t$, where A is state-transition matrix, $\mathbf{w}_t$ is $6-$dimensional zero mean Gaussian i.i.d. noise, independent of state and with covariance matrix U which specifies the extent of noise.

B. Shape and color cues

The contour cues can be very useful to represent the appearance of the tracked objects with distinctive silhouette when a model of the shape can be learned off-line and then adapted over time. The shape of the head is one of the most easily recognizable human parts and can be reasonable well approximated by an ellipse. In our approach the oval shape of each head candidate is verified using the sum of intensity gradients along the ellipse's boundary.

When the contour information is poor or is temporary unavailable, color information can be very useful alternative to extract the tracked object. Color as a cue is computationally inexpensive. A color histogram including spatial information can be extracted on the basis of a 2-dimensional kernel centered on the target [3]. In order to assign smaller weights to the color of pixels that are further away from the center of the kernel a nonnegative and monotonic decreasing function $k : [0, \infty) \rightarrow R$ can be utilized [3]. The probability of particular histogram bin u at location $\mathbf{x} = \{x, y\}$ is determined by the following formula: $d_{\mathbf{x}}^{(u)} = C_r \sum_{l=1}^{L} k\left(\left\|\frac{\mathbf{x}-\mathbf{x}_l}{r}\right\|^2\right) \delta\left[h(\mathbf{x}_l) - u\right]$ where $\mathbf{x}_l$ are pixel locations, L is the number of pixels in the considered kernel, r is the radius of the kernel, δ is the Kronecker delta function, and the function $h : R^2 \rightarrow \{1, ..., K\}$ associates the bin number. The normalization factor C_r ensures that $\sum_{u=1}^{K} d_{\mathbf{x}}^{(u)} = 1$. It can be precalculated for the utilized kernel and assumed values of r [3]. The 2-dimensional Gaussian kernels have been prepared off-line and then stored in lookup tables for the future use. The color representation of the target has been extracted by quantizing the ellipse's interior colors into K bins and extracting the weighted histogram. To make the histogram representation of the tracked head less sensitive to lighting conditions the V component obtained the 4-bin representation while the remaining components of the HSV color space have been represented by 8 bins.

To compare the histogram Q representing the tracked face to histogram I obtained from a particle configuration we utilized the metric $\sqrt{1 - \rho(I, Q)}$ that is derived from Bhattacharyya coefficient $\rho(I, Q) = \sum_{u=1}^{K} \sqrt{I^{(u)} Q^{(u)}}$.

The work [3] showed that the used metric is invariant to the scale of the target and therefore is superior to other measures. Using the Bhattacharyya coefficient we defined the color observation model as $p(\mathbf{z}^C \mid \mathbf{x}) = (\sqrt{2\pi}\sigma)^{-1} e^{-\frac{1-\rho}{2\sigma^2}}$. Thanks to such weighting we favor head candidates whose color distributions are similar to the distribution of the tracked head. The second ingredient of the observation model reflecting the edge strength along the elliptical head boundary has been weighted in a similar manner: $p(\mathbf{z}^G \mid \mathbf{x}) = (\sqrt{2\pi}\sigma)^{-1} e^{-\frac{1-\phi_g}{2\sigma^2}}$, where ϕ_g denotes the normalized gradient along the ellipse's boundary.

C. Appearance model of the face

The appearance model of the face holds a sample of RGB color components for each pixel in the ellipse's interior and uses this sample to estimate the probability density function of pixel's color in the current frame. If $C_{\mathbf{x}}^{(j)} = \{c_1^{(j)}, c_2^{(j)}, ..., c_M^{(j)}\}$ is a recent sample of a color component j for a pixel at location $\mathbf{x}$, the probability density function that the component j of this pixel will have value $c_t^{(j)}$ at time t can be non-parametrically estimated using the kernel K_h as $p(c_t^{(j)}) = \frac{1}{M}\sum_{i=1}^{M} K_h(c_t^{(j)} - c_i^{(j)})$. For Gaussian kernel $K_h = \mathcal{N}(0, \Sigma)$ and a given sample $C_{\mathbf{x}}^{(j)}$ from a distribution with density $p(c^{(j)})$, where $\Sigma = (\sigma^{(j)})^2$ is the kernel bandwidth, an estimate of this density at $c^{(j)}$ can be calculated as follows: $p(c^{(j)}) = \frac{1}{M}\frac{1}{\sqrt{2\pi}\sigma^{(j)}}\sum_{i=1}^{M}\exp\left(-\frac{1}{2}\frac{(c^{(j)} - c_i^{(j)})^2}{(\sigma^{(j)})^2}\right)$. The kernel bandwidth should express the local variation in the value. In our approach the bandwidth assumes different values over the face region and changes over time. Assuming that $c^{(j)}$ is distributed according to $\mathcal{N}(\mu, \sigma^2)$, the distribution of deviation for each consecutive pair $c_i^{(j)} - c_{i+1}^{(j)}$ is $\mathcal{N}(0, 2\sigma^2)$. Hence the standard deviation can be estimated as $\sigma^{(j)} = \frac{1}{0.68\sqrt{2}}\frac{1}{(M-1)}\sum_{i=1}^{M-1}|c_i^{(j)} - c_{i+1}^{(j)}|$ [10]. Assuming independence between the pixel color components in the considered RGB color space, the probability estimate that the examined pixel belongs to face can be expressed as follows: $p(c) = \frac{1}{M}\sum_{i=1}^{M}\prod_{j=1}^{3}\frac{1}{\sqrt{2\pi}\sigma^{(j)}}\exp\left(-\frac{1}{2}\frac{(c^{(j)} - c_i^{(j)})^2}{(\sigma^{(j)})^2}\right)$. The pixel is considered as belonging to face area if $p(c) < th$, where th is a global threshold over the ellipse's interior. We assumed the following observation model $p(\mathbf{z}^A \mid \mathbf{x}) = \frac{1}{L}\sum_{l=1}^{L}\delta\left[p(c_{\mathbf{x}_l}) < th\right]$, where L is the number of pixels in the interior of the model ellipse of fixed size.

D. Probabilistic integration of cues

The aim of probabilistic multi-cue integration is to enhance visual cues that are more reliable in the current context and to suppress less reliable ones. Assuming that the observations are conditionally independent given the state we obtain the equation $p(\mathbf{z}_t \mid \mathbf{x}_t) = p(\mathbf{z}_t^G \mid \mathbf{x}_t) \cdot p(\mathbf{z}_t^C \mid \mathbf{x}_t) \cdot p(\mathbf{z}_t^A \mid \mathbf{x}_t)$ which allows us to accomplish the probabilistic integration of cues. To achieve this we compute at each time t the L2 norm-based distances $D_t^{(d)}$, between the individual cue's centroids and the centroid obtained by integrating the likelihood from utilized cues [11]. The reliability factors of the cues $\alpha_t^{(d)}$ are then calculated on the basis of the following leaking integrator: $\xi\dot{\alpha}_t^{(d)} = \eta_t^{(d)} - \alpha_t^{(d)}$, where ξ denotes a factor that determines the adaptation rate and $\eta_t^{(d)} = 0.5 * (\tanh(-aD_t^{(d)}) + b)$. In the experiments we set $a = 0.3$ and $b = 3$. Using the reliability factors the observation likelihood has been determined as follows: $p(\mathbf{z}_t \mid \mathbf{x}_t) = [p(\mathbf{z}_t^G \mid \mathbf{x}_t)]^{\alpha_t^{(1)}} \cdot [p(\mathbf{z}_t^C \mid \mathbf{x}_t)]^{\alpha_t^{(2)}} \cdot [p(\mathbf{z}_t^A \mid \mathbf{x}_t)]^{\alpha_t^{(3)}}$, where $0 \le \alpha_t^{(d)} \le 1$.

E. Face detection-based proposal for particle filter

The face detection algorithm can be utilized to form a proposal distribution for the particle filter in order to direct the particles towards most probable locations of the objects of interest. The employed face finder is based on object detection algorithm described in work [13]. Taking the location and the size of the window containing the face we construct a Gaussian distribution $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{z}_t)$ in order to reflect the face position in the proposal distribution. The formula describing the proposal distribution has the following form: $q(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{z}_t) = \beta p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{z}_t) + (1-\beta)$ $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$. The parameter β is dynamically set to zero if no face has been detected during tracking. In such a situation the particle filter takes the form of the CONDENSATION [7].

F. Adaptation

The color histogram has been updated over time in the following manner: $Q_t^{(u)} = (1-\gamma)Q_{t-1}^{(u)} + \gamma I_t^{(u)}$, where γ is an accommodation rate, Q_{t-1} is the previous histogram representing the tracked face, I_t denotes the histogram from the interior of ellipse determined by the state estimate, whereas $u = 1, ..., K$. In the appearance model the oldest sample element is discarded and a new one is added as it becomes available. The extent of noise that is added to $x, \dot{x}, y$, and $\dot{y}$ during the prediction stage has been determined on the basis of the reliability factors $\alpha_t^{(d)}$. Since fewer particles are needed for noise with small variance the number of particles has also been adjusted with respect to $\alpha_t^{(d)}$.

IV. QUANTIZATION STRATEGY

The H.264/AVC reference software JM 2.2 was modified to include our face tracking system. Different quantization parameters have been used for macroblocks belonging to the face region and macroblocks that belong to the background. The quantization parameters have been first optimized for a target face quality in terms of PSNR. The quantization step Q_i for the i^{th} face macroblock was determined in the following manner [9]: $Q_i = \sqrt{12\frac{D_{target}}{E}\frac{N_f\sigma_i}{\sum_{i=1}^{N_f}\sigma_i}}$, where N_f is the number of face macroblocks, $D_{target} = 255^2 \cdot 10^{-\frac{PSNR}{10}}$, σ_i is the standard deviation, and E is the ratio between the true distortion and a model distortion. The true distortion is known at the end of the frame encoding and the model distortion is given as: $D_{model} = \frac{1}{N_f}\sum_{i=1}^{N_f}\frac{Q_i^2}{12}$.

A redistribution of available bits has been dictated by the given bit budget constraint. The foreground quality was overspecified if the bit budget for the frame was less than the total number of bits allocated for the face encoding. In such a situation the background was encoded with the worst possible quality. To minimize the subjective quality degradation along detected face borders the PSNR of the background macroblocks has been determined as a function of their distance to the foreground macroblocks. To satisfy the bit-rate as closely as possible a rate of gradual decrease of the background PSNR has been accommodated over time.

V. Experiments

To test the proposed method of face tracking we performed various experiments. We utilized the MissAm sequence as our first test set. The face has been found in all images and the system tracked the face in the whole sequence. The Carphone sequence is challenging for face detection algorithm because it contains a non-upright face expressing several emotions. The detection algorithm which has been trained to detect only upright faces has detected about 50% face-regions in the discussed sequence. To improve the detection performance each image in the sequence has additionally been rotated about 20 deg and then verified in respect of face presence. For algorithm operating on both rotated and non-rotated images only 4 of 381 face-regions have not been detected and only one false detection has been observed. The track of the face has been correctly kept in all frames of the sequence. In the Foreman test sequence all faces has been detected in the first 100 frames of sequence. In the first 200 frames of the sequence the face has not been detected in 40 frames and no false detection has been observed. The system tracked the face in all images of the sequence. It is worth to note that using the above mentioned test sequences the face can be successfully tracked using the CONDENSATION algorithm and the following observation model: $p(\mathbf{z}_t \mid \mathbf{x}_t) = [p(\mathbf{z}_t^G \mid \mathbf{x}_t)]^{\alpha_t^{(1)}} \cdot [p(\mathbf{z}_t^C \mid \mathbf{x}_t)]^{\alpha_t^{(2)}}$. The observation model presented in section 3.4 has acknowledged its usefulness in experiments consisting in tracking a face with an active camera in front of wooden doors or furniture. The reliability factor $\alpha_t^{(3)}$ is particularly helpful in determining the extent of noise as well as the quantity of particles.

The H.264 encoder was used to perform intraframe coding at a target bitrate of 45 kbps. At a somewhat smaller bitrate the Y-PSNR of the facial region in the Foreman sequence was improved by 2.4 dB, whereas the image quality of the non-facial region was degraded by only 1.6 dB. Fig. 1 presents the frame #100 of the Foreman sequence that was encoded with unmodified and modified H.264 encoder.

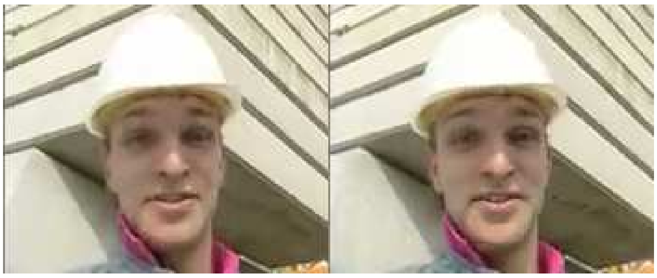

Fig. 1. Foreman frame #100 encoded with unmodified H.264 (left) and modified H.264 (right)

VI. Conclusions

In this paper a face detection and tracking algorithm has been used to provide the region of interest for the H.264 encoder. The algorithm operates within the syntax of H.264 and no decoder modifications are required. At the same bitrate a higher quality of the facial region was obtained. The appearance model of face improves the reliability of tracking and can be useful for control the number of particles.

References

[1] ITU-T and ISO/IEC JTC1, "Advanced video coding for generic audiovisual services," ITU-T Recommendation H.264-ISO/IEC 14496-10 AVC, 2003.

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

[3] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," In *Proc. IEEE Conf. on Comp. Vision and Pattern Rec.*, pp. 142-149, 2000.

[4] S. Daly, K. Matthews, and J. Ribas-Corbera, "Face-based visually optimized image sequence coding," In *Proc. IEEE Int. Conf. on Image Proc.*, pp. 443-447, 1998.

[5] A. Doucet, S. Godsill, and Ch. Andrieu, "On sequential Monte Carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, pp. 197-208, 2000.

[6] A. Eleftheriadis, and A. Jacquin, "Automatic face location detection and tracking for model-assisted coding of teleconferencing at low bit-rates," *Signal Processing: Image Communication*, vol. 7, no. 3, pp. 231-248, 1995.

[7] M. Isard, and A. Blake, "Contour tracking by stochastic propagation of conditional density," *European Conf. on Computer Vision*, pp. 343-356, 1996.

[8] C. W. Lin, Y. J. Chang, and Y. C. Chen, "Low-complexity face-assisted video coding," In *Proc. IEEE Int. Conf. on Image Proc.*, pp. 207-210, 2000.

[9] C. Perra, M. Pinna, and D. D. Giusto, "H.263+ rate control at fixed objective quality," In *Proc. of PV2000 Packet Video Workshop*, Italy, May 2000.

[10] D. W. Scott, "Multivariate Density Estimation," New York, Wiley, 1992.

[11] J. Triesch, and Ch. von der Malsburg, "Democratic integration: Self-organized integration of adaptive cues," *Neural Computation*, vol. 13, pp. 2049-2074, 2001.

[12] B. Yang, L. Wu, and A. Waibel, "Focus of attention: Towards low bitrate tele-conferencing," In *Proc. IEEE Int. Conf. on Image Proc.*, vol. 2, pp. 97-100, 1996.

[13] P. Viola, and M. Jones, "Rapid object detection using a boosted cascade of simple features," In *Proc IEEE Conf. on Comp. Vision and Pattern Rec.*, pp. 511-518, 2001.

[14] H. Wang, and S-F. Chang, "A highly efficient system for automatic face region detection in MPEG video," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 7, no. 4, pp. 615-628, 1997.

[15] T. Wiegand, G. J. Sulliwan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560-576, 2003.