AN AUTOMATIC METHOD FOR UNEQUAL AND OMNI–DIRECTIONAL ANISOTROPIC DIFFUSION FILTERING OF VIDEO SEQUENCES

Adriana Dumitraş and Jim Normile

Apple Computer
2 Infinite Loop, MS:302–3KS, Cupertino, CA 95014, United States
Email: adrianad@ieee.org, jnormile@apple.com

ABSTRACT

In this paper we propose a method for automatic, unequal and omni–directional anisotropic diffusion filtering of video sequences. Our method, which consists of automatic foreground/background discrimination using the color characteristics of the human face and skin, unequal filtering of the determined regions and then composition of the filtered regions into the final frames, is applied prior to encoding the video sequences. In addition to yielding higher visual quality of the decoded sequences than that of the decoded sequences with equal filtering over the entire frame, our method is robust with respect to skin tone and lighting variations, has low complexity and requires no calibration. We illustrate the main advantages of our method using videoconferencing sequences.

1. INTRODUCTION

Achieving bit rate reduction in the compressed video streams depends on the selection of a video coding method and the characteristics of the video content. To condition the video content for better compression, filtering methods can be employed [1]. Spatial filtering methods achieve tradeoffs between the amount of detail that is being removed and the visual quality of the filtered sequence. Typically, removing sufficient details for compression yields noticeable distortion in the filtered sequences in the form of excessive smoothing of flat areas and blurring of the edges. Such artifacts may be particularly disturbing when they affect regions–of–interest (ROIs) such as faces in videoconferencing applications.

A solution to reduce the impact of spatial filtering artifacts is to process the ROIs differently than the rest of the video frames. This idea has been used in foveated vision works [2], which exploited the ability of the human visual system to focus only on a main region at a time, hence making possible stronger compression of the other frame areas. For such solutions to be viable in real–time videoconferencing applications, automatic identification of the regions to be filtered and a low complexity of the filtering method are required. The former requirement has been addressed by tracking of visual features (face shape, relative position of eyes, nose and mouth) and tracking of gaze. However, more than one frame is required initially to determine correctly the ROI placement, or a tedious user–dependent calibration of the gaze tracker is needed. Moreover, the complexity associated to tracking the ROIs may be quite high. The latter requirement has been addressed by various implementations of spatial filters that typically achieve a trade–off in terms of picture quality and processing speed.

To address these limitations, in this work we propose an automatic and low complexity method for region–of–interest (ROI) filtering in video sequences. In our method, the foreground and the background regions are determined automatically using the color characteristics of the face and skin, then they are filtered differently. Filtering is performed so as to reduce the details while preserving sharp edges in the foreground, and smoothing the areas in the background using omni–directional anisotropic diffusion. In addition to yielding higher visual quality of the decoded sequences with unequal pre–filtering than that of the decoded sequences with equal filtering over the entire frame, being robust with respect to skin tone and lighting variations, our method has low complexity and requires no calibration.

The rest of the paper is organized as follows. Section 2 presents basic ideas related to face/skin identification in video sequences and traditional anisotropic filtering. Our proposed method in presented in detail in Section 3. The relationship of the proposed work with other spatial filtering methods is discussed in the same section. Experimental results and conclusions are included in Sections 4 and 5, respectively.

2. BACKGROUND

In this section we discuss briefly face/skin identification in video sequences using color information and spatial filtering by anisotropic diffusion. Face identification using color information has been performed with the objective of determining accurately the location of the human face in videoconferencing sequences [3, 4], as well as other scenes. Methods that define precisely chroma maps for the human skin and take advantage of its particular color characteristics across human races, and methods that make use of histograms have been employed. The former category of methods is more precise than the latter, however, details related to modeling of individual color features have not been disclosed [4].

Among the spatial filtering methods, diffusion filtering has the important property of generating a scale space via a partial differential equation. In the scale space, analysis of object boundaries and other information at the correct resolution where they are most visible can be performed [5]. Two classes of diffusion filtering methods have been proposed and they perform isotropic and anisotropic diffusion, respectively. Isotropic diffusion methods make use of the isotropic diffusion equation (the heat equation) given by $\frac{\partial \mathcal{I}(x,\, y,\, t)}{\partial t} = div(\nabla \mathcal{I})$ [6] . This equation is applied using the original (noisy) image $\mathcal{I}(x, y, 0)$ as the initial condition, where $\mathcal{I}(x, y, 0) : \Re^2 \longrightarrow \Re^+$ is an image in the continuous domain, (x, y) specifies a pixel's spatial position, t is a time parameter and $\nabla \mathcal{I}$ is the image gradient. Modifying the image according to this isotropic diffusion equation is equivalent to filtering the image using a Gaussian filter [6]. Unfortunately, the distortion introduced in highly textured areas is visible immediately and accounts for an important part of the picture impairments identified by the viewers at first sight.

The anisotropic diffusion methods make use of the equation proposed in [5, 7], which replaces the classical isotropic diffusion

equation above with $\frac{\partial \mathcal{I}(x,\, y,\, t)}{\partial t} = div\,[g(\parallel \nabla \mathcal{I} \parallel)\nabla \mathcal{I}]$, where $\|\nabla \mathcal{I}\|$ is the magnitude of the image gradient and $g(\|\nabla\|)$ is an "edge stopping" function. This function satisfies the condition $g(x) \longrightarrow 0$ when $x \longrightarrow \infty$ such that the diffusion is stopped across the edges.

Because of their ability to reduce details in images without impairing the subjective quality, anisotropic diffusion filters have been applied to smoothing of pictures, segmentation and object tracking. In image and video compression, anisotropic diffusion filters have been employed as a pre–processing step prior to still image (JPEG) compression [8]. They have also been applied as post–filters to the decoded frames in order to remove blocking and ringing artifacts [9, 10].

3. PROPOSED METHOD

Our proposed method for unequal and omni–directions filtering of video sequences consists of four steps, which are illustrated in Fig. 1. An example of the outcome of each of these steps is illustrated in Fig. 2. First, we identify the face/skin regions–of–interest using their color characteristics. Second, we construct a corresponding binary mask M_{fg} for the foreground region, as well as a complementary mask M_{bg} for the background region. Third, we filter the foreground and background regions differently using anisotropic diffusion in eight directions and the parameter sets P_{fg} and P_{bg}, respectively. Fourth, we combine the filtered regions into a final picture. These steps are described next.

We perform identification of the foreground ROI using the chromatic characteristics of the human face/skin by identifying all of the pixel locations $(i,\, j)$, that satisfy the condition $U_{low} \leq U(i,\, j) \leq U_{high}$ AND $V_{low} \leq V(i,\, j) \leq V_{high}$, where U and V are the chroma planes, and $1 \leq i \leq M$, $1 \leq j \leq N$, with $M \times N$ the picture size. Then, we define the binary mask of the foreground M_{fg} such that it contains pixel values equal to 1 within the largest bounding box of the pixels identified above, and 0 otherwise. The binary mask of the background M_{bg} is, in turn, the complement of M_{fg}.

Next, we apply unequal anisotropic diffusion filtering of the foreground and background ROIs as follows. We extend the traditional Perona–Malik anisotropic diffusion filter [7] given in a discrete form by

$$\mathcal{I}(x,\, y,\, t+1) \;=\; \mathcal{I}(x,\, y,\, t) + \lambda \sum_{p \in \eta(x,\, y)} g(\nabla \mathcal{I}_p(x,\, y, t))\; \nabla \mathcal{I}_p(x,\, y\, t) \tag{1}$$

where $\mathcal{I}(x,\, y,\, t)$ is a discrete image, $(x,\, y)$ denotes the pixel position in a discrete, 2–D grid and t denotes discrete time steps (iterations). The scalar constant $\lambda \in \Re^{+}$ determines the rate of diffusion, $\eta(x,\, y)$ represents the spatial neighborhood of the pixel located at position $(x,\, y)$ [7, 6]. Traditional Perona–Malik anisotropic filter performs diffusion in four directions (north, south, east and west) with respect to a pixel location (x, y). By extending the filter to perform diffusion in eight directions (north, south, east, west, north–east, south–east, south–west, north–west), the above equations becomes in two dimensions:

$$\mathcal{I}(x,\, y,\, t+1) \;=\; \mathcal{I}(x,\, y,\, t) + \lambda\,[\sum_{N,S,E,W} c_m(x,\, y,\, t)\; \nabla \mathcal{I}_m(x,\, y,\, t) + \beta \sum_{NE,SE,SW,NW} c_n(x,\, y,\, t)\; \nabla \mathcal{I}_n(x,\, y,\, t)\,] \tag{2}$$

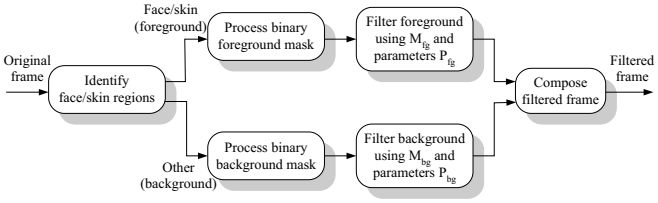


Fig. 1. Block diagram of the proposed method. Notations M_{fg}, M_{bg} stand for the foreground and background masks, respectively. Notations P_{fg}, P_{bg} stand for the parameter sets for filtering the foreground and background regions, respectively.

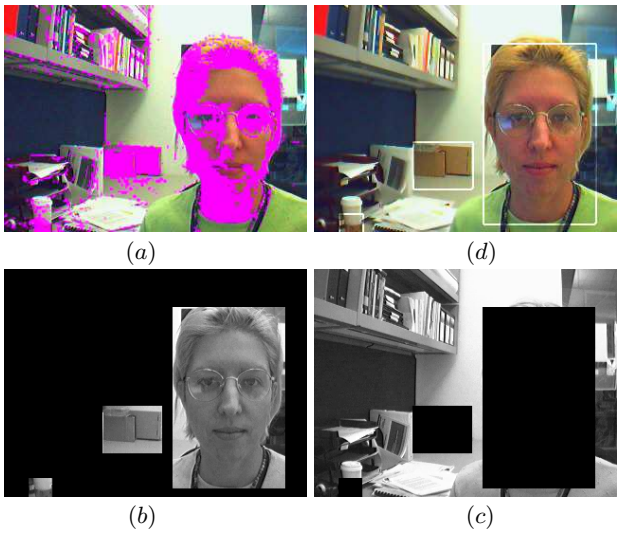


Fig. 2. Example of the processing steps of our proposed method: (a) identified face/skin regions (in magenta), (b) foreground region, (c) background region and (d) final filtered regions (in white boxes) in the frame that is being sent to the encoder.

where the subscripts correspond to the eight directions of diffusion with respect to the pixel location (x, y), m is one of the traditional (N, S, E, W) directions, n is one of the diagonal (NE, SE, SW, NW) directions, λ is less or equal to $\frac{1}{|\eta(x,\, y)|}$, where $|\eta(x,\, y)|$ is the number of neighbors (eight, except at the image boundaries, therefore $\lambda \leq \frac{1}{8}$). Notations c_m, c_n stand for diffusion coefficients in traditional (N, S, E, W) directions indexed by m and diagonal directions (NE, SE, SW, NW) directions indexed by n. Each diffusion coefficient is typically an edge stopping function of $\nabla \mathcal{I}(x,\, y,\, t)$ in the corresponding direction, i.e.,

$$\begin{aligned} c_m(x,\, y,\, t) &= g(\nabla \mathcal{I}_m(x,\, y,\, t)) \\ c_n(x,\, y,\, t) &= g(\nabla \mathcal{I}_n(x,\, y,\, t)) \end{aligned} \tag{3}$$

Because the distance between a pixel location (x, y) and any of its diagonal neighbors is larger than the distance between the distance between a pixel location and its vertical/horizontal neighbors, we scale the diagonal pixel differences[1] by a factor β, which is a function of the frame dimensions M, N.

We also employ the approximation of the image gradient in a selected direction of [7] and given by (4). For instance, in the

[1]This is similar to making use of the idea of square and non–square pixels within images.

"northern" direction the gradient can be computed as the difference given by (5).

$$\nabla \mathcal{I}_p(x,\, y,\, t) \;=\; \mathcal{I}_p(x,\, y,\, t) \;-\; \mathcal{I}(x,\, y,\, t), \;\; p \in \eta(x,\, y) \quad (4)$$

$$\nabla \mathcal{I}_N(x,\, y) \;=\; \mathcal{I}(x,\, y+1,\, t) \;-\; \mathcal{I}(x,\, y,\, t) \quad (5)$$

For the face/skin (foreground) regions, we select the set of parameters $\mathcal{P}_{fg} = \{k_{fg}, \lambda_{fg}\}$ and the edge stopping function given by (6). For all of the other background regions, we select the background set of parameters $\mathcal{P}_{bg} = \{k_{bg}, \lambda_{bg}\}$ and the edge stopping function given by (7) [7]. Notations λ_{fg}, λ_{bg} stand for the values of the parameter that determines the rate of diffusion in eq. (2) in the foreground and background, respectively. Finally, we assemble the filtered ROIs into a final picture.

$$g(x,\, y,\, t) \;=\; \frac{1}{1 + \left(\frac{\nabla \mathcal{I}(x,\, y,\, t)}{k_{bg}} \right)^2} \quad (6)$$

$$g(x,\, y,\, t) \;=\; \exp\left[- \left(\frac{\nabla \mathcal{I}(x,\, y,\, t)}{k_{fg}} \right)^2 \right] \quad (7)$$

Our method features distinctive features as well as similarities with previous works. Similarly to [4] which evaluates chroma ranges of the human face/skin, we make use of the fact that, for all human races, these ranges are consistently the same. As opposed to the work in [4], we determine chroma ranges that are robust for many videoconferencing sequences. Moreover, we do not make use of histograms and do not require exact face segmentation, since we make use of the maximum bounding box associated with the face region. Consequently, the complexity of our face/skin detection stage is much lower than that of [4].

By using the edge stopping functions given by eqs. (7) and (6), our method inherits the advantages of the traditional anisotropic filter, namely inexpensive implementation and rapid smoothing. By extending the filter to perform omni–directional diffusion, we improve the effectiveness of the filtering stage. Unlike the works in [9] and [10], which make use of anisotropic filtering as a post–processing step to remove blocking and ringing artifacts, respectively, and similarly to the works in [8], we perform anisotropic filtering in the pre–processing stage. However, the latter work has focused on still image compression using the traditional version of the anisotropic diffusion filter, whereas we apply omni–directional and unequal anisotropic diffusion filtering to video sequences.

4. EXPERIMENTAL RESULTS

We have evaluated the bit rate reduction and visual quality obtained using our method and H.263 coded videoconferencing sequences. The video test set employed in our experiments consists of head–and–shoulder sequences acquired during real–time videoconferencing using with various cameras such as Orange iBOT and Logitech Notebook Pro. The frames are represented in YUV format. The frame rate is equal to 15 and 30 frames/sec (fps). The frame size is equal to 352×288 pixels per frame (CIF format). The ranges of values for U and V have been determined in this work experimentally, using twenty video sequences that included high quality movies and sequences acquired using various cameras. We have determined that $U_{low} = 75$, $U_{high} = 130$, $V_{low} = 130$ and $V_{high} = 160$ yield good face/skin identification for numerous test sequences. Our selection of the chromatic range for face/skin regions is in line with that tested in [4], however, the actual ranges of chroma values are different. We evaluate the effectiveness of the proposed method using the visual quality of the decoded sequences with and without filtering by our method, compared at the same bit rate.

We have applied our filter with the parameter values $k = 50$, $\lambda = 0.2$ to filtering of sequences acquired using the above mentioned cameras. Next, the movie sequences have been encoded using an H.263 codec with rate control at 150 kbits/s. The subjective quality of the decoded pictures is quite different, as Fig. 3 illustrates. More specifically, our proposed method yields a sharper appearance of the face ROI in Fig. 3 (b) as compared to that of the same ROI in Fig. 3 (c) when equal filtering of the entire frame has been applied prior to encoding. For better comparison, enlarged sections of these frames are illustrated in Fig. 4. As illustrated in Fig. 5, the proposed method is robust with respect to large variations of the illumination conditions. We have also verified that the face/skin identification procedure is robust with respect to large variations of the skin tone. More specifically, skin tones corresponding to African–American and Asian–American subjects are also identified correctly.

5. CONCLUSIONS

We have presented a method for unequal anisotropic diffusion filtering of video sequences that yields very good subjective quality of the decoded frames in a H.263–based videoconferencing system. Our method employs automatic identification of the face and skin regions using their color characteristics, followed by anisotropic diffusion filtering. The latter is here applied in eight directions and makes use of different edge stopping functions and parameters for the skin/face (foreground) regions and other (background) regions, respectively. We have shown that our proposed unequal and omni–directional anisotropic diffusion filtering yields much better visual quality of the decoded pictures in a H.263–based videoconferencing system than equal filtering in fewer directions. Moreover, our method is robust with respect to variations of illumination and skin tone, and has low complexity.

6. REFERENCES

[1] Barry G. Haskell, Atul Puri, and Arun N. Netravali, *Digital Video: An Introduction to MPEG–2*, Chapman and Hall, USA, 1997.

[2] Sanghoon Lee, Marios S. Patichis, and Alan Conrad Bovik, "Foveated video compression with optimal rate control," *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 977–992, July 2001.

[3] Christophe Garcia and Georgios Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Trans. on Multimedia*, vol. 1, no. 1, pp. 264–277, Sept. 1999.

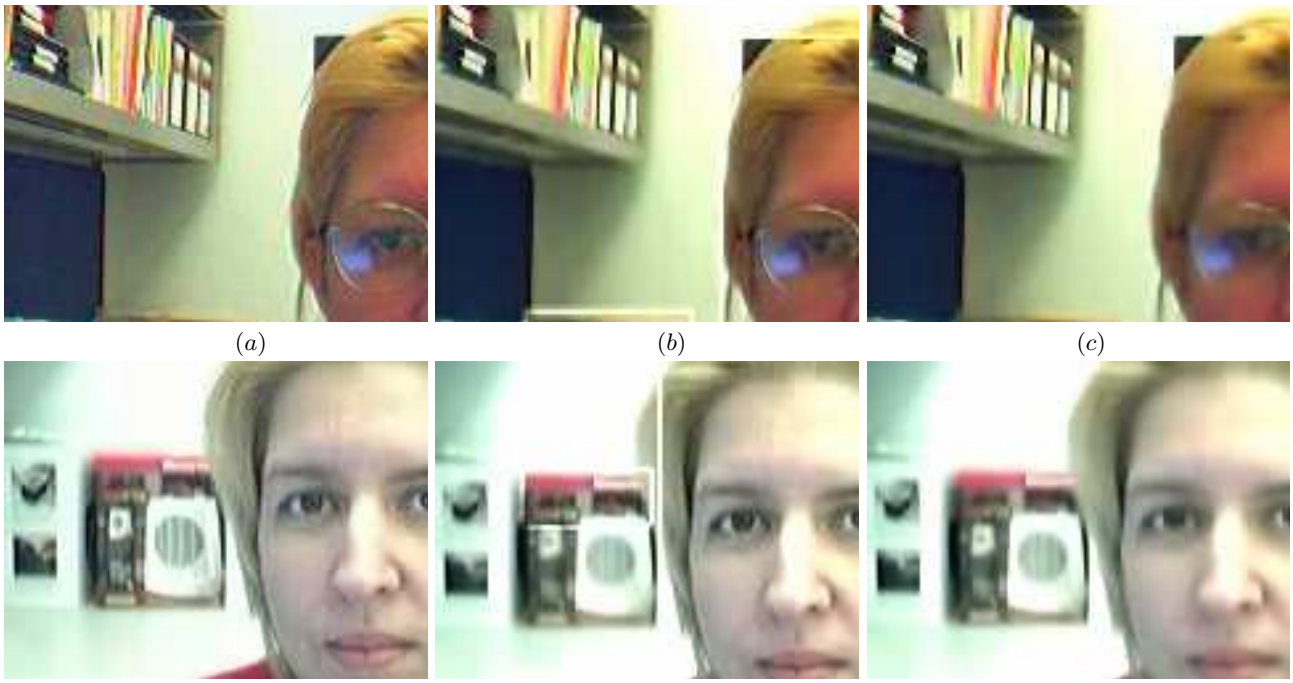
[4] Douglas Chai and King N. Ngan, "Face segmentation using skin–color map in videophone applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 551–564, June 1999.

[5] Pietro Perona, "Anisotropic diffusion processes in early vision," in *Proc. of the IEEE Multidimensional Signal Processing Workshop*, Pacific Grove, CA, USA, 1989, pp. 68–71.

[6] Michael J. Black, Guillermo Shapiro, David H. Marimont, and David Heeger, "Robust anisotropic diffusion," *IEEE Transactions on Image Processing*, vol. 7, no. 3, pp. 421–432, Mar. 1998.

(a) (b) (c)

Fig. 3. (a) Original frame acquired using a Logitech Notebook Pro camera, (b) frame that has been filtered using our proposed method and then coded at 150 kbits/s using a H.263 video codec, and (c) frame that has been filtered using *equal* anisotropic diffusion and then coded at 150 kbits/s using a H.263 video codec.



(a) (b) (c)

(d) (e) (f)

Fig. 4. Details of the original and decoded frames after coding using H.263 at 150 kbps: (a) original (Orange iBOT) frame from Fig. 2, (b) coded after filtering using our method, (c) coded after filtering using equal anisotropic filtering, (d) original (Logitech Notebook Pro) frame in Fig. 3, (e) coded after filtering using our method, and (f) coded after filtering using equal anisotropic filtering.

[7] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, July 1990.

[8] I. Kopilovic and T. Sziranyi, "Nonlinear scale-selection for image compression improvement obtained by perceptual distortion criteria," in *Proc. of the International Conference on Image Analysis and Processing*, Venice, Italy, 1999, pp. 197–202.

[9] Yang Seungjoon and Yu-Hen Hu, "Coding artifacts removal using biased anisotropic diffusion," in *Proceedings of IEEE International Conference on Image Processing*, Santa Barbara, CA, USA, 1997, vol. 2, pp. 346–349.

[10] P.W. Devaney, D.C. Gnanaprakasam, and T.J. Leacock, "Post-filter for removing ringing artifacts of DCT coding, US patent #5819035, Oct. 6, 1998.," .

(a) (b)

Fig. 5. Robustness of the face/skin detection under low lighting conditions: (a) original frame and (b) frame with identified face/skin (in magenta).