THE ARBITRARILY SHAPED TRANSFORM OF SEGMENTED MOTION FIELD FOR A PSEUDO OBJECT-ORIENTED VERY LOW BIT-RATE VIDEO CODING SYSTEM

Chung-Wei Ku, You-Ming Chiu, Liang-Gee Chen, and Yung-Pin Lee

DSP/IC Design Lab., Department of Electrical Engineering
National Taiwan University, Taipei, Taiwan, R.O.C.
Email: {william,lgchen}@video.ee.ntu.edu.tw
URL: http://video.ee.ntu.edu.tw/william

ABSTRACT

The arbitrarily shaped transform (AST) of the segmented motion field generated by modified optical flow algorithm (MOFA) is suggested in this paper. Because the transform kernels are dependent on the shape of the coded region, information about the contours of the objects must be transmitted. To achieve data compression, coefficient selection function (CSF) is defined for AST coefficients reduction. Since shape information costs lots of data amount, a rectangle approximation method is also developed. According to the simulation results, the proposed coding method is efficient and its performance is much better than the traditional block matching algorithm no matter in terms of PSNR or bit-rate. In addition, the proposed algorithms are integrated together to build a very low bit-rate video coding system. From some primary simulations of the system, the proposed method can guarantee 34 dB PSNR for most cases at around 10 Kbps; Furthermore, the segmentation of objects is potential for object scalability.

1. INTRODUCTION

Although H.261 has defined a standard for videophone or videoconferencing with $p\times64$ Kbps, lower bit-rate is expected in order to utilize current general switched telephone network (GSTN). According to the standard of modern modem (V.34), transmission rate is defined as 28.8 Kbps. To meet the above requirements, as early as May 1991 the "Moving Picture Expert Group" (MPEG) raised the issue of audio-visual standard targeted at the bit-rate of 4.8-64 Kbps, with the motivation being the limited channel bandwidth and limited storage capacity. These efforts was approved in July 1993 with the MPEG-4 nickname and the title "Very Low Bit-Rate Coding of Moving Pictures and Associated Audio" [1]. Recently, ITU-T also announced a draft about "Video Coding for Low Bit-Rate Communication" or H.263 [2], which is an extension of H.261 standard for very low bit-rate visual telephone. It is still a block based algorithm basically with some MPEG-1 optimizations. However, since the goal of MPEG-4 is to integrate telecommunication, computer, and multimedia applications, some quite different approaches about MPEG-4 are also developed rapidly. For example, model-based coding [3] and object-oriented coding algorithms [4]. We developed a modified optical flow based motion estimation algorithm to generate a homogeneous and meaningful motion field [7]. In order to encode these motion vectors efficiently and build a prototype of a very low bit-rate video coding system, arbitrarily shaped transform and coutour description are proposed and applied to the segmented motion field in this paper. Besides, a pseudo object-oriented very low bit-rate video coding system which is based upon the above techniques is presented. The simulation results reveal the proposed system is very suitable for very low bit-rate video coding applications, or even content-based applications.

2. ARBITRARILY SHAPED TRANSFORM

The general form of the two dimensional unitary transform is

$$F(i,j) = \sum_{x,y} f(x,y)k(i,j,x,y), \tag{1}$$

and its inverse transform is

$$f(x,y) = \sum_{i,j} F(i,j)k(i,j,x,y). \tag{2}$$

$f(x,y)$ indicates the image data in spatial domain and $F(i,j)$ indicates the coefficients in frequency domain. k is the kernel function of the transformation. x,y are spatial indices and i,j are frequency indices. To meet the unitary restriction, k must satisfy the following property:

$$\sum_{x,y} k(i,j,x,y)k(m,n,x,y) = 0, \quad for\ i \neq m\ or\ j \neq n.$$

$$\sum_{x,y} k(i,j,x,y)k(i,j,x,y) = 1, \quad for\ i = m\ and\ j = n.$$

In most developed standards such as H.261, MPEG-1 and MPEG-2, DCT plays an important role for spatial redundancy reduction. However, if there are edges in the transformed block, the performance of DCT degrades obviously. To avoid the above drawbacks, an object-oriented *arbitrarily shaped transform* (AST) is suggested in literature [6]. It can be used for any shape of homogeneous regions rather than the square block for general DCT. Data compression is achieved by transmitting the coefficients in frequency domain. On the other hand, since the input data distribute in arbitrary shape, the kernel functions of the transform are dependent on the distribution. In this paper, polynomial based functions are selected in our system for the feasibility of programming. Because the basis should satisfy the orthonormal requirements for arbitrarily shaped conditions, an orthonormalization process for those polynomial basis must be executed, such as Gram-Schmidt process; more details can be found in [6]. In brief, the procedures for arbitrarily shaped transform includes:

1. For the region which will be transformed, find its circumscribed rectangle.

0-7803-3073-0/96/$5.00 ©1996 IEEE

2. According to the shape of the region, orthonormalize the polynormal basis.
3. Apply the transformation to the input region according to the orthonormalized basis to obtain the coefficients.

The basis to be orthogonalized are $\{1, x, y, xy, x^2, y^2 \ldots\}$. It should be noticed that the basis are orthonormalized in both encoder and decoder. In addition, the shape information must be transmitted because the kernels are orthonormalized according to the shape of the coded region; the decoder can recover the kernels from the shape information then apply inverse arbitrarily shaped transform (IAST) to those coefficients to reconstruct the original input.

For the reason of data compression, only parts of the low-ordered transformation coefficients are computed and transmitted. Besides, as the sizes and shapes of different regions vary widely, it is unfair to calculate a fixed number of coefficients for any kinds of regions. That is, some regions need more coefficients while some others need less. Our approach is to determine how many and which coefficients are necessary before applying forward transform. Generally speaking, the number of critical coefficients is dependent on the size and homogeneity of the region. If the region is considerably homogeneous, we assume two heuristic rules to select the coefficients. The first is: "For large regions, more coefficients are needed to reach the same performance as that for small regions". It is a reasonable assumption because large regions contain more information than small regions statistically. The second rule is: "The increasing of necessary coefficients in the first rule should be fewer than the increasing of region size". To interpret it, consider the case of DCT. Suppose block A is four times as large as block B. When the homogeneities in both blocks are comparable, the number of critical coefficients for block A should be less than four times of that for block B. This phenomenon also can be observed in the case of AST. Figure 1 shows the realization of the coefficient selection scheme. As Figure 1 (a) shows, given the region we can find its circumscribed rectangle with width p and height q. The frequency domain after AST is shown in Figure 1 (b). We compute only the effective coefficients in the shaded area which is with two sides, $f(p)$ and $f(q)$. According to the above two heuristics, the *coefficient selection function* (CSF) f should satisfy the following properties:

1. f is an increasing function.
2. The deviation of f (f') is a decreasing function.

In our system, $f(x) = \sqrt{x}$ is selected as the CSF, which satisfies the above properties and its complexity is acceptable.

3. SEGMENTATION FOR AST

Before applying AST, the motion field is segmented into several homogeneous regions. In fact, the operation of segmentation is equivalent to clustering similar motion vectors around. The similarity of the neighboring pixels can be evaluated from some local informations. That is, if the difference between a motion vector and one of its neighbor is large, the two vectors should be segmented into different regions; otherwise they will be grouped into the same region. It is analogous that a drop of color ink drips down the canvas and floods over the region with similar texture, where the similarity is defined as the gradient of the segmented field. Since there are two components in a motion vector (i.e. x-component and y-component), both the gradients in x and y field are considered for the segmentation. If both of the gradients are larger than a threshold, the considered vectors will not be grouped together. The above segmentation method is very suitable for later AST operation because each segmented motion field keeps its homogeneity. In addition, the number of regions are quite few for real conversation applications because there are only one or two persons on screen in general. An example explaining the above ideas is shown is Figure 2. Since the kernels are dependent on the shape information, it is unnecessary to transmit the data about kernels; instead, the shape of the coded region is used. In most cases, there is only one outer boundary enclosing the region. However, if one region covers part of another, as shown in Figure 3 region A has both inner and outer boundaries, it is redundant to encode the common boundary between A and B twice. Instead, we regard A as a hollow-free region. As long as A is decoded before B, where region B locates will be reconstructed consequently. Since region A is encoded like those regions without holes, the pixels in the hole (the region enclosed by dashed curve in Figure 3) must be filled with some value. It is reasonable to give them the mean value of the object. Suppose there are m pixels in A and n pixels in the hole, then the mean value of A is $\frac{1}{m}\sum_{i=1}^{m} E_i$. If we insert the hole with this value, the new mean is the same:

$$\frac{1}{n+m}\left(\sum_{i=1}^{m} E_i + n\left(\frac{1}{m}\sum_{i=1}^{m} E_i\right)\right) = \frac{1}{m}\sum_{i=1}^{m} E_i.$$

Therefore, the first coefficient after AST for the new hollow-free object will be equal to that for original A.

The encoding of region contour will be inefficient if error-free coding scheme is applied, such as chain code. It causes that large amount of the encoded data is about contour. Instead, the shape is approximated by some descriptions before appling AST. There are many methods for curve approximations, and they must be able to resolve the situations of overlaps and "no-mans-land" holes. Fortunately, since the error is not noticeable for human eyes, it is not really a serious problem. In addition, the possibility of the above phenomenon is quite few in general conversation applications. In our system, the shape description is approximated by rectangle. A heuristic algorithm is developed to find those knot points. These knot points are linked as a rectangle to approximate the region as Figure 4 displays. The segmented and approximated motion field is then applied with AST. Because the motion vectors vary little in spatial domain, AST can be executed in "grid unit" instead of "pixel unit". That is, the segmented regions is determined in subsampled motion field. In our experiments, the subsampling factor is 4 for both x-axis and y-axis; the computation complexity is reduced greatly while the performance remains excellent. Before appling AST in the subsampled motion field, both x and y fields are with the same shape described by the rectangle knot points. The polynomial based kernels of AST is orthonormalized according to the approximated shape. Two sets of AST coefficients are computed and transmitted to the channel/storage resource. Since the information about motion and shape are both preserved, the proposed methods are very suitable for very low bit-rate coding or motion recognition.

4. EXPERIMENTAL RESULTS

The performance of the proposed AST and rectangle description for motion field is evaluated by several monochromatic head and shoulders sequences, which are with QCIF

format and 10 frames/sec frame rate. In order to simplify the evaluation, the residual error will not propagate to the later frames in our simulation. If the data of two successive frames are subtracted directly without any motion compensation method, the PSNR would be the lower bound with respect to motion estimation/motion compensation. The bit-rate of frame difference is zero because nothing is transmitted. On the other hand, if the the motion vectors generated by MOFA is used for motion compensation directly, its PSNR would be the upper bound with respect to AST of the motion field. However, the bit-rate will be very high because the motion vector of each pixel must be transmitted. The result of the proposed AST coding is between the two bounds because of the loss by encoding motion vectors. The simulation results reveal that the proposed AST coding of the motion field can guarantee 35 dB PSNR picture quality at least. In addition, the rectangle description method works pretty well with almost neglectable loss in performance. To make a clear comparison, the average performance of several sequences is listed in Table 1. In this table, *FS* indicates full-search block matching algorithm with block size 8×8 and search range -4 to $+4$. Remember that *MOFA* is the upper bound while *frame diff.* is the lower bound. For the sequence "*Caisy*" and "*Elsa*", the proposed method outperforms *FS*. For "*Miss America*" and "*Jian*", the performance is about 1 dB lower than FS. It is obvious the performance of the proposed AST with rectangle description is almost equal to full-search block matching algorithm, but the data amount is reduced. The data amount analysis is displayed in Table 2. The average chain code length and AST coefficient numbers are all listed. The full-search block matching algorithm needs 6 bits for each motion vector (-4 to +4), and there are $\frac{176 \times 144}{8 \times 8} = 396$ blocks in each frame. The total data for one frame is $396 \times 6 = 2376$ bits. On the other hand, the proposed method needs to transmit one shape information and two sets of AST coefficients. For chain code description each chain costs 3 bits; the AST coefficients are with 10 bits word length in our system. Therefore the average data amount for one frame is about one quarter of the amount of FS with comparable performance. In addition, if the rectangle description scheme is applied, the data amount is reduced further more. In our simulation, the x and y parts of the knot points are coded with 6 bits (-32 to +32) each. As a result, extra 20% saving of AST data is obtained but the performance is still preserved. All the materials are variable length coded to remove the statistical redundancy. To decode the bitstream, In inter mode, a motion field is generated by inverse arbitrarily shaped transform (IAST) from the shape (i.e. the orthonormal kernels are generated from the shape), then current picture is reconstructed by motion compensation with the preceding decoded frame and the motion field. According to some primary simulation results of our system, the proposed methods are very useful for very low bit-rate video coding applications. In general, 10 Kbps bit-rate can be expected.

5. CONCLUSION

We proposed a modified optical flow based motion estimation algorithm (MOFA) [7]. Compared with other motion estimation algorithms, its performance is the best for both subjective view and PSNR. In order to transmit the motion field efficiently, an arbitrarily shaped transform (AST) process is designed in this paper. To achieve the goal of data compression, coefficient selection function (CSF) is carefully defined and rectangle approximation of shape is developed in our system. According to the simulation results, AST of the motion field is efficient and the performance is much better than the traditional block matching algorithm no matter in terms of PSNR or bit-rate.

In the proposed pseudo object-oriented system, motion estimation algorithm is executed by MOFA to remove the temporal redundancy in video sequence. The generated motion field is subsampled and segmented into regions; these regions are applied with AST to remove the spatial redundancy in motion field. For conversation application, usually there is only single object on screen and most of the contents on screen are well compensated by the above combination; this "pseudo object" is called motion compensated object (MCO). The coding of motion off object (MFO) is under development. Since the patterns appear on screen is less restricted in our system, we believe the proposed system is very suitable and practical for videophone or videoconferencing.

To optimize the system, the investigation about fast AST is studied because it is the most time-consuming part in our system. Currently the encoding time is the most serious problem and the real-time software decoder is under programming. Besides, for some detail operations of eyes and mouth, several fine compensation methods will be appended to advance the acceptance of picture [8]. In addition to very low bit-rate applications, the object scalability is implied in our region segmentation and more efforts will be spent on this topic.

REFERENCES

[1] ISO/IEC JTC1/SC29/WG11 MPEG 92/699, *Project Description for Very Low Bitrate A/V Coding*, 5, 1992.

[2] ITU-T Study Group 15, Working Party 15/1, "Draft ITU-T Recommendation H.263 - Video coding for low bitrate communication", *Document-LBC-95-251*, Jul. 1995.

[3] K. Aizawa, H. Harashima and T. Saito, "Model-based analysis-synthesis image coding (MBASIC) system for a person's face", *Signal Processing: Image Communication*, vol. 1, pp. 139-152, 1989.

[4] H. Shiller and M. Hötter, "Investigations on color coding in an object-oriented analysis-synthesis coder", *Signal Processing: Image Communication*, vol. 5, pp. 319-326, 1993.

[5] N. Ahmed and K.R. Rao, *Orthogonal Transform for Digital Signal Processing*, New York: Springer-Verlag, 1975.

[6] M. Gilge, "Coding of arbitrarily shaped image segments based on a generalized orthogonal transform", *Signal Processing: Image Communication*, vol 1, pp. 153-180, 1989.

[7] C.-W. Ku, Y.-M. Chiu, L.-G. Chen, and Y.-P. Lee, "Building a pseudo object-oriented very low bit-rate video coding system from a modified optical flow motion estimation algorithm" , *to appear in Proc. ICASSP'96, Atlanta.*

[8] C.-W. Ku, L.-G. Chen, Y.-M. Chiu, and Y.-P. Lee, "A pseudo object-oriented very low bit-rate video coding system with cache VQ for detail compensation", *submitted to IEEE International Conference on Image Processing, Laussane*, Sep. 1996.

Table 1. Average performance for several approaches.

Seq. name	*frame diff.*	*MOFA*	*MOFA+ AST*	*MOFA+AST+RECT.*	*FS*
Miss America	32.772	41.175	37.635	37.389	38.389
Caisy	37.734	43.624	41.616	41.059	39.995
Elsa	35.713	42.662	39.476	39.226	39.455
Jian	32.403	40.856	37.392	36.976	38.582

Table 2. Average data amount of the proposed method.

Name	Length	Chain	Knot No.	Coeff. No.	AST (bit/frame)	AST+RECT. (bit/frame)
Miss America	47	98.17	12.61	18.39	662.31	519.12
Caisy	32	66.50	9.00	13.43	468.10	376.6
Elsa	51	89.92	11.00	17.46	618.96	481.2
Jian	26	85.92	9.20	18.64	630.56	483.20

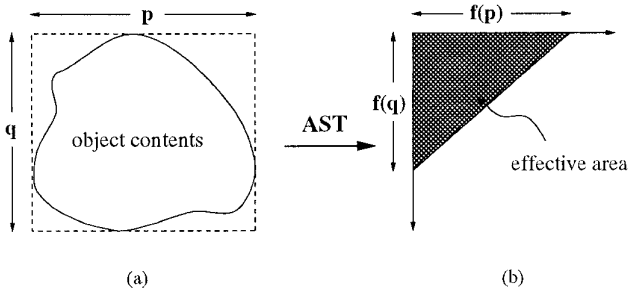

Figure 1. Coefficient selection scheme: (a) spatial domain, (b) frequency domain.

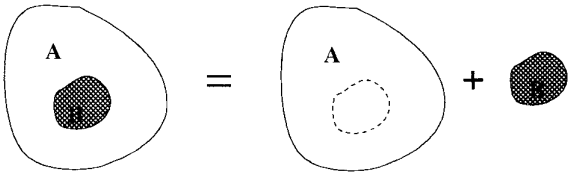

Figure 3. Shape coding for enclosed regions.

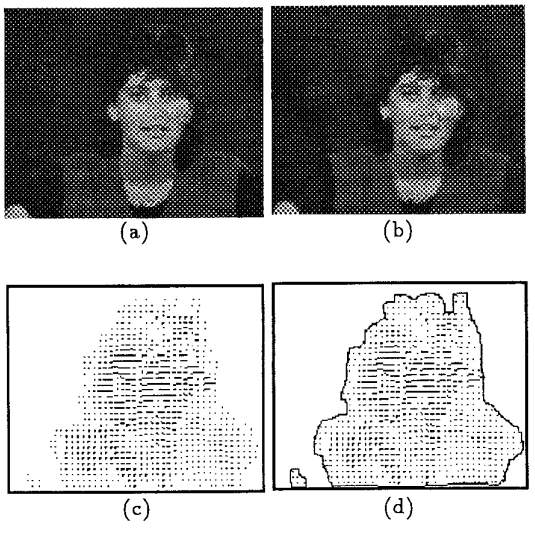

Figure 2. An example: (a) previous frame, (b) current frame, (c) motion field, and (d) segmented regions.

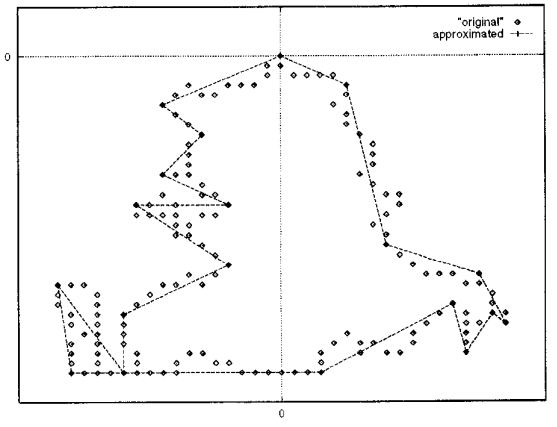

Figure 4. Rectangle description for shape approximation.