On Active Camera Control and Camera Motion Recovery with Foveate Wavelet Transform

Jie Wei, *Member*, *IEEE Computer Society*, and
Ze-Nian Li, *Member*, *IEEE*

Abstract—In this paper, a new variable resolution technique–Foveate Wavelet Transform (FWT) is proposed to represent digital images in an effort to efficiently represent visual data. Compared to existing variable resolution techniques, the strength of the proposed scheme encompasses its linearity preservation, orientation selectivity, and flexibility while supporting interesting behaviors resembling the animate vision system. The linearity preservation of the FWT is due to the fact that only low and/or high-pass filterings are carried out in different regions of an image in the transform. The orientation selectivity indicates the fact that details along the horizontal, vertical, and diagonal directions are readily available in the FWT representation. The flexibility of this new representation technique is witnessed by the readiness of its extensions to represent foveae of different number, shape, and locations. To demonstrate the efficacy of the FWT, two applications are presented. First, an FWT-based active camera control scheme is developed, where the computer can move a camera to track the moving object in the scene. Second, an FWT-based method purporting to recover pan/tilt/zoom camera movements from video clips is developed. Experiments of these two applications have shown encouraging performances.

Index Terms—Active vision, wavelet transform, variable resolution techniques, gaze control, object tracking, motion detection.

1 Introduction

In most vision applications, due to the great amount of visual information involved in relatively high resolution pictures taken by cameras, the demand on computation, storage, and communication is prohibitive. As described in [2], in order to cope with the great amount of information posed by nature, the animate vision system (AVS) has two areas: *fovea area* (FA) and *periphery area* (PA). The FA, generally a small portion of the entire view, provides detailed information; whereas the PA, which covers much wider viewing angles, offers the background information and incurs little processing load. With these two different areas, two types of eye movements carry out the routine tasks: *catching* and *holding*. The former, in the form of saccades, are resulted from a shift of attention; whereas the latter, in the form of smooth pursuits, are employed to hold an object and follow it when it moves. The latency of these responding movements is minimized due to the existence of the foveal and peripheral structures of the AVS. In this manner, the AVS works efficiently and effectively, which sheds a light on some tracks for possible digital image and video representations.

Motivated by the AVS, an ensemble of methods referred to as variable resolution (VR) techniques have been developed. Three of them are introduced briefly below. The images before and after the VR transforms are denoted as retinal and cortical images, respectively.

1. *Log-Polar transforms:* This group of transforms is due to the biological research conducted by Schwartz [13], where it is concluded that the mapping between the retinal image and cortical image for human eyes can be approximated by a logarithmic function. The Log-Polar transform, one of the most frequently used techniques [11], [6], is formulated as follows: $z' = \log z$ if $|z| > r_f$, otherwise $z' = z$. Where $z = re^{i\theta}$ is the complex variable corresponding to the representation of a point (x, y) in the Descartes plane, thus $x = r\cos\theta$ and $y = r\sin\theta$; r_f is the radius of the fovea in the shape of a circle.

   The Log-Polar transform offers a close emulation to the AVS fovea-peripheral structure. By exploiting the polar coordinates, it simplifies centric scaling and rotation [22], because these transformations now become shift operations in the $\log r$ and θ dimensions, respectively. As shown by Sandini and Dario [11], the scaling and centric rotational invariances of the Log-Polar transform make it a useful tool for object recognition. In a more recent project, Sandini et al. [12] successfully developed a videophone system for elderly and disabled people, where a camera mounted on a pan-tilt unit with a Log-Polar CMOS sensor delivered video data in an efficient space-variant manner. In spite of its advantages, there is however, a major drawback of the Log-Polar transform. In general, image patterns of linear structures and translating movements are distorted into streamlines of log-sine curves [22] which adversely complicates the analysis of these common problems in computer vision.

   In order to combat the irregular cortical images after the logarithmic transform and the multiple foveae, Basu and Wiebe [23], [24] proposed several updated Log-Polar mappings: The moving fovea, the stretched and Cartesian variable resolution transforms.
2. *Reciprocal wedge transform (RWT):* The RWT transform [16] is defined by the following mappings from a rectangular-shaped retinal image into a wedge-shaped cortical image: $\mathrm{w} \simeq \mathrm{Tz}$, where $\mathrm{w} = (u, v, 1)^T$ and $\mathrm{z} = (x, y, 1)^T$ are the corresponding homogeneous coordinates of a point in the cortical and retinal images, respectively. The 3×3 matrix T is the cross-diagonal unit matrix. With this matrix notation, some commonplace operations such as translation, rotation, and scaling can be formulated as the multiplication of corresponding matrices, whereby a much easier image manipulation is fostered. The RWT encompasses valuable characteristics such as simplicity and linear structure preservation. While its shortcoming includes the irregular shape of the resultant cortical image and the limited representation power—only the streak-shaped animate visual systems can be represented.
3. *Multiresolution-based transform:* The central idea shared by all multiresolution VR techniques is to represent the fovea with higher resolution and the periphery with lower resolution in a pyramidal representation. A hierarchical architecture for the representation and multiprocessing of foveal images, referred to as foveal polygon, has been developed in [1]. The merits of this type of techniques are that: 1) The mappings involved are linear; 2) the shape of the cortical image is regular, and 3) it is easy to implement. However, the major problem is its inability to emulate the AVS due to the

- *J. Wei is with the Department of Computer Science, City College of City University of New York, Convent Ave. at 138th St., New York, NY 10031. E-mail: wei@cs.ccny.cuny.edu.*
- *Z.-N. Li is with the School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada, V5A 1S6. E-mail: li@cs.sfu.ca.*

Manuscript received 24 July 1998; revised 9 Nov. 2000; accepted 26 Feb. 2001. Recommended for acceptance by H. Christensen.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 107860.

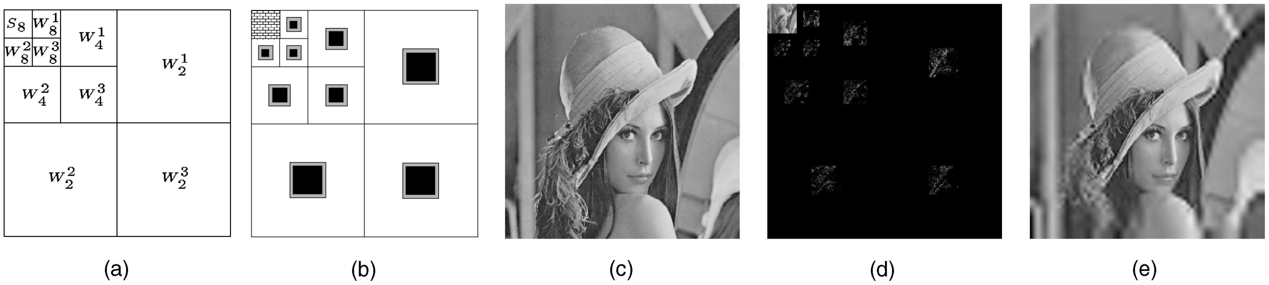


Fig. 1. (a) Discrete wavelet transform. (b) FWT mask ϖ: brick-patterned area–preserved S_8, deeply shaded area–OFA, lightly shaded area–PPA, and blank area—DPA. (c) Original image $\mathcal{I}$. (d) FWT $\mathcal{F}(\mathcal{I})$. (e) Reconstructed image.

presence of the resolution discontinuity across neighboring acuity levels, which also complicates the visual processing across those levels. Based on multiple spatial resolution representation, Klarquist and Bovik [8] developed a foveated vergent stereo active vision system FOVEA for the purpose of recovering 3D scenes. Chang and Yap [3] achieved foveation by manipulating wavelet coefficients in the context of visualization. Without consideration of the perfect reconstruction of the FA and smooth resolution transition, this scheme is still an implementation of the multiresolution scheme.

In this paper, the Foveate Wavelet transform (FWT), a novel wavelet-based VR technique to represent an image with spatially variable resolution, is proposed. The FWT provides various merits such as its linearity preservation, orientation selectivity, and high flexibility while exhibiting desirable visual effects reflecting the VR result. The linearity preservation of the FWT is due to the fact that only low and/or high-pass filterings are carried out in different regions of an image during the transform. The orientation selectivity indicates the fact that details along the horizontal, vertical, and diagonal directions are readily available in the FWT representation. The flexibility of this new representation technique is witnessed by the readiness of its extensions to represent foveae of different number, shape, and locations.

Two applications of the FWT are presented in this paper. First, by analogy to the eye movements of the AVS, we develop a scheme carrying out active camera control based on the FWT, which is the essential task of exploratory and purposive vision. As a second important application of the FWT, a method to recover pan/tilt/zoom camera motion from videos is introduced, wherein a motion detection algorithm is first applied to FWT-based frames to acquire dense motion fields. Next, pan/tilt/zoom camera motions and their combinations are recovered by inspecting the behaviors of the motion fields.

These two applications are in the research areas of object tracking and motion detection which have received extensive attentions in computer vision and image processing communities. State-of-the-art object tracking methods usually exploit prior high-level knowledge [5] and a multitude of visual cues in order to track objects of interest, e.g., salient image features by Smith and Brady in ASSET [14] and moving corner cluster by Reid and Murray [10]. Some of the existing motion detection techniques include:

1. MRF-MAP framework: The Markov Random Field (MRF) is used to model the visual constraints and the Maximum a Posteriori (MAP) is searched to obtain the most likely configuration [25].
2. Simultaneous multiple motion detection: Regions of different motion are obtained simultaneously through a clustering process [17].
3. Dominant motion detection: Within each iteration a single region of dominant motion is segmented out. This process is then iterated after the previous region of dominant motion is removed [7].

Object tracking and motion detection can find a broad spectrum of civil and military applications in areas such as intelligent video surveillance, robot control, video processing, and moving object extraction.

2 Foveate Wavelet Transform

2.1 Discrete Wavelet Transform

Localization in both frequency and time/space domains is the greatest advantage of the discrete wavelet transform (DWT) over Fourier transform-based methods, e.g., Discrete Fourier Transform, which are only localized in the frequency domain. The spatial localization indicates that after the wavelet transform the coefficients in a certain position at the wavelet subimages correspond to the details of different frequencies in the corresponding spatial location. Whereas the frequency localization indicates the fact that each subimage of the wavelet transform corresponds to the information of a single frequency and orientation for the original image.

The core of the DWT is two filters, one is of low-pass or averaging nature, denoted as h, the other is of high-pass or differential nature, denoted as g. To obtain the DWT, first one applies h and g, respectively, in the vertical direction and decimates the results by a factor of two. W_{vl} and W_{vh} are obtained. Next, h and g are applied to W_{vl} and W_{vh} in the horizontal direction. After decimation, four subimages, W_{ll}, W_{lh}, W_{hl}, and W_{hh}, are derived. They are denoted as S_2, W_2^1, W_2^2, and W_2^3, which represents a coarser version, the vertical details, the horizontal details, and the diagonal details of the original image respectively. This process can be further repeated on S_2. In Fig. 1, a three-level DWT is depicted. To reconstruct the image from the DWT, the reconstruction filters are used, which are h and g or the duals h^* and g^* for orthonormal and biorthonormal wavelets, respectively. Wavelets whose h and g filters are of the length m and n, respectively, are denoted as m/n wavelets. Biorthonormal wavelets are usually used in image processing because there is no nontrivial orthonormal linear phase filters with the perfect reconstruction property [9].

2.2 The Foveate Wavelet Transform

Following the notation of Fig. 1, the FWT is introduced using an FWT mask ϖ which is the union of subimages ϖ_j^is (with the same size as W_j^is) and ϖ_8 which corresponds to S_8.

Definition. *For image $\mathcal{I}$, its FWT $\mathcal{F}(\mathcal{I})$ is $\varpi \cdot DWT(\mathcal{I})$, where $DWT(\mathcal{I})$ is the DWT of $\mathcal{I}$, "$\cdot$" is the pixel wise multiplication of corresponding images.*

Assuming that m/n wavelets are applied, the FA is a square of the size s, which is one-fourth of the side of the original image located at the center. Assume the origin of the coordinate system of each subimage is at its center and $M = Max\{m, n\}$. The *OFA*, *Original FA*, corresponds to the spatial area of the original FA. The *PPA*, *Preserved PA*, refers to the part of PA whose coefficients in the detail wavelet subimages are preserved. The *DPA*, *Discarded PA*, indicates the part of PA whose detail coefficients are discarded.

$$(x, y) \in \begin{cases} \text{OFA} & \text{if both } |x| \text{ and } |y| \in [0, \frac{s}{2^j}], \\ \text{PPA} & \text{if both } |x| \text{ and } |y| \in [0, \frac{s}{2^j} + \frac{3}{4}M] \text{ and } (x, y) \notin OFA, \\ \text{DPA} & \text{otherwise.} \end{cases} \tag{1}$$

The values of the components of the FWT mask ϖ are then assigned as follows: 1) $\varpi_8(x, y) = 1$, for all (x, y) in S_8 and 2) within ϖ_j^i: $\varpi_j^i(x, y)$ is 1 if $(x, y) \in OFA \bigcup PPA$, 0 otherwise.

The reconstructed image of the FWT is the inverse wavelet transform starting from $\mathcal{F}(\mathcal{I})$. The above definition of the FWT indicates that all the detail information in the OFA and PPA in the DWT representation of a given image is preserved, while those in the DPA are discarded. The proposed FWT and the reconstructed image for Lenna is depicted in Fig. 1. Due to the preceding formulation of the FWT, the following advantages are present:

1. Direct image analysis in the transform domain: This is due to localization of the wavelet transform in both the spatial and frequency domains. Because of the frequency localization, details in the FA of a certain frequency can be extracted from the corresponding detail subimages. For instance, if the information with low frequencies is of utmost interest, then our detection algorithm needs only be conducted on the detail subimages of higher levels, such as W_8^is and W_4^is as depicted in Fig. 1. Due to the spatial localization, the feature detection algorithms which use spatially localized operators can be applied directly on the portions of the FWT. Therefore, the most important image analysis operations such as feature detection and feature match can be applied directly in the transform domain, which is impossible for most VR techniques. In [20], various direct image analyses have been applied to the FWT successfully.
2. Orientation selectivity: As described previously, for the discrete wavelet transform, detail subimages W_j^1s, W_j^2s, and W_j^3's correspond to the vertical, horizontal, and diagonal details, respectively. This type of orientation selectivity is unique to the DWT and is preserved in the PPA/OFA's. Hence information with respect to these three directions in PPA/OFA's is readily available in the FWT representation and can lend themselves to an efficient information retrieval. In [19], for the purpose of detecting stereo disparities, where of major concern is the vertical edges, the disparity search can be conducted only on the PPA/OFAs in W_j^1's, a mere one-third of the original FWT representations, thus resulting in a significant computational economy.

   Apparently, other VR techniques provide their own "orientation selectivity." For example, for Log-Polar transform, this occurs along the polar dimensions of r and θ. While this kind of "orientation selectivity" is beneficial to certain object-centered operations such as looming/zooming [12], it is less applicable to most common operations and movements which can be best captured and described in the Cartesian image space where the orientation selectivity of the FWT applies as shown in the preceding stereo example [19].
3. Smooth transition across FA and PA: In general, the closer the position of a pixel is from the FA, the higher resolution is exhibited there. This is achieved by the following two factors:

   a. *The presence of the PPAs in the detail subimages.* In the FWT, the PPAs at all levels are of the same width. Recall that more decimations occurred at higher levels of the wavelet subimages representing less details, thus the effective coverage of the PPA's increases when going up the wavelet subimage hierarchy. This is similar to what is in the multiresolution scheme, except the PPA masks are now used and defined in the wavelet transform domain. As a result, the lower frequency coefficients in the PPAs at higher levels have their impact on larger spatial areas in the reconstructed images.

   b. *The shape and nature of the reconstruction filters.* Generally, the magnitudes of the nonzero items $h^*(k)$ and $g^*(k)$ decrease as the magnitude of the index k increases. In addition, since the lengths of the two filters generally employed in image processing tasks are larger than two, their application thus overlaps, which indicates that the value of one wavelet coefficient makes contributions to the reconstructed values of those positions which are neighbors of its corresponding spatial area.

The presence of the PPAs can also ensure the perfect reconstruction of the FA. Due to the overlapness of the reconstruction filters as previously described, in order to render the FA as a perfect reconstruction area, the preserved area in detail subimages should have extra areas around the OFA which spatially corresponds to the FA. Because for the inverse wavelet filters g^* and h^* the magnitude δ of the nonzero items with the largest index number is extremely small, the contribution of the coefficients with distance greater than $\lceil M/2 \rceil$ in the inverse wavelet transform, which is a sequence of convolutions, is negligible. Therefore, the minimal value of w to achieve a perfect reconstruction of FA is $\frac{M}{2}$. For the m/n wavelets used in the image processing community, the values of m and n are usually less than 10. If one sticks to the $M/2$, the width of the transition area is too narrow and sharp to simulate the AVS. Hence, we adopt $w = 0.75M$ as the width of the PPAs in the FWT representation in order to result in a wider and smoother transition area.

From the above definition of the FWT, it bears some resemblance with other multiresolution schemes in achieving the spatially variable resolution. However, the major difference between these two techniques lies in the fact that the FWT manipulates the frequency/spatial details and the VR is achieved by the intentional removal of frequency/spatial details. Whereas, most multiresolution techniques realize variable resolution by working entirely in the spatial domains. In [18], we show that, in addition to the reduced space consumption and inherent powers of wavelet transform such as the space and frequency localization and orientation selectivity, the FWT has the valuable feature of emulating the AVS more closely than its multiresolution counterparts, i.e., the resolution realized from the reconstructed image of the FWT representation exhibits smoother transition from the foveal area to the peripheral area than that of the multiresolution representation.

With the FWT, unlike other VR techniques, no disparity exists with different shape, number, and positions of the FA. The

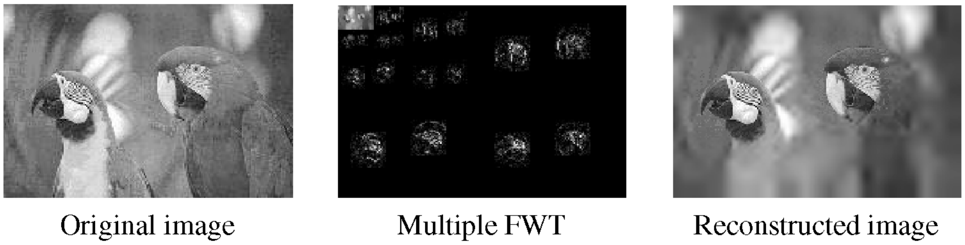


Fig. 2. Illustration of multiple FAs.

multiple foveae [24] and streak-shaped fovea [16] can be achieved with ease under the framework of the FWT, i.e., by employing more OFAs/PPAs and a streak-shaped FA, respectively. As described in [24], there are cases where more than one FA is present for animate vision systems. It is the same case for images or videos, where more than one portion may be of utmost interest. It is rather unwieldy for other VR techniques to cope with these cases. Wiebe and Basu discovered two elegant yet tricky strategies, namely, cooperative and competitive methods, to attain a representation of multiple foveae [24]. An image with two-FAs is shown in Fig. 2, where the two FAs are around the heads of the two birds. Therefore, the flexibility of the FWT is clearly demonstrated.

3 Active Camera Control

3.1 FWT-Based Active Control Algorithm

Equipped with the concepts of FA and PA, the FWT can be employed to actively govern the imaging parameters of a camera, which is referred to as *gaze control* in [15]. The catching movements now indicate a rapid panning/tilting of the camera (saccade) in order to capture certain object of significant motion in the periphery (DPA or PPA). The holding movements, in the form of minor panning/tilting or zoom, are meant to better track moving objects inside the FA. The previous two camera movements are denoted as *gaze change* and *gaze stabilization*, respectively, which are two primary categories of gaze control [15]. In our implementation, the camera control scheme is based solely on the motion of surrounding objects in the FWT representation, which can be approximated by the Foveate Potential Moving Area (FPMA). The process of creating the FPMA is indeed the difference method carried out on the FWT representation. It is obtained by the following three steps: 1) differencing and thresholding adjacent FWT representations in the highest level, 2) propagating the nonzero areas in the upper level to lower levels according to the inter-band spatial relationships of the DWT, and 3) conducting component labeling based on 8-connectivity, components whose area is smaller than a given threshold are deleted. The resulting FPMA is a collection of blobs of considerable size. Therefore, the centroid of each FPMA blob can be viewed roughly as the center of a moving object, thereby sophisticated motion detection process is bypassed to effect a real-time response. Based on the FWT, the following algorithm purporting to control the camera motion actively is proposed, where C is the centroid of one or many FPMA blobs, f_{opt} is the object to be employed to draw the attention of the camera centered at O_{new}.

```
PROCEDURE FPMA-FWT
  1. Preprocessing:
     Compute the area A of each FPMA blob $f$ in terms of the
        original resolution;
  2. Pan/tilt control:
     IF there exist $f$'s whose area A(f) is greater than $\delta_1$
        set $O_{new}$ to $C(f_{opt})$ where $f_{opt} = argMax_f\{A(f)\}$;
     ELSE IF there exist f's whose A(f) s larger than $\delta_2$
        set $O_{new}$ to $C(f_{opt})$, where $f_{opt} = \sum_F f$
           and $F = \{f : A(f) > \delta_2\}$;
     ELSE no change is made to $O_{new}$ and set $f_{opt}$ to the
           area with the largest area;
        $f_{opt} = argMax_f\{A(f) : f \bigcap OFA \neq \emptyset\}$;
     point the center of the camera to $O_{new}$ through pan/tilt;
  3. Zoom control:
     IF $O_{new}$ is inside OFA: zoom-out if $A(f_{opt}) \geq \delta_3$,
        zoom-in if $A(f_{opt}) \in (\delta_4, \delta_5)$
END FPMA-FWT
```

The δs ($\delta_1 > \delta_2, \delta_4 > \delta_5$) are prescribed thresholds. In Step 2, if there exists an FPMA blob whose area exceeds δ_1, the centroid of the one with the most area is opted as O_{new}; in cases where no FPMA has an area greater than δ_1, the one-object constraint is assumed, because of this the FPMA's with area larger than δ_2 are viewed as from the same object, thus the centroid of the ensemble of these blobs claims O_{new}; otherwise, O_{new} remains unchanged. In Step 3, the zoom is only issued if the O_{new} lies in the OFA. When the corresponding area is greater than δ_3, a zoom-out is to be issued; while if the area lies in between δ_4 and δ_5, a zoom-in is invoked. In our implementation, by trial and error, the values opted for δ_1, δ_2, δ_3, δ_4, δ_5 are $A_0/32$, $A_0/128$, $A_0/64$, $A_0/128$, $A_0/200$, respectively, A_0 is the area of the original frame.

3.2 Experimental Results

The proposed active camera control scheme FPMA-FWT has been implemented and tested extensively using a pan-tilt-zoom camera on a NOMAD mobile robot. All computations are performed by the on-board computer. Due to the FWT representation, where the PA undergoes a low-pass process, the camera is more discriminative to motions inside the FA while minor motions and noises outside the FA are ignored. To illustrate, if a motion of small magnitude, say seven, is present in the PA of the current frame, then in its FWT representation, no difference from the FWT of the reference frame can be found, thereby no FPMA is labeled for this motion. On the other hand, the camera is also set to be alert to relatively significant apparent motions in the periphery so as to invoke saccadic movements for gaze change.

As for the efficiency of the FPMA-FWT, two factors are involved: 1) The new VR technique FWT reduces the search space to about one-eighth of the original size, and 2) the camera motion is determined by the FPMA, which is generated through a process of masking and component labeling. In our experiments, the average time for determining each camera movement is 0.47 second.

In order to have a comparison between the presence and absence of the FWT as the representation, an algorithm based on the PMA of the original frame is also implemented, denoted as PMA-ORIGINAL. Another algorithm is implemented using the PMA based on the whole wavelet transform of the original frame, denoted as PMA-WAVELET. Notice that for these two schemes, except without the OFA, PPA and DPA concepts, the PMA

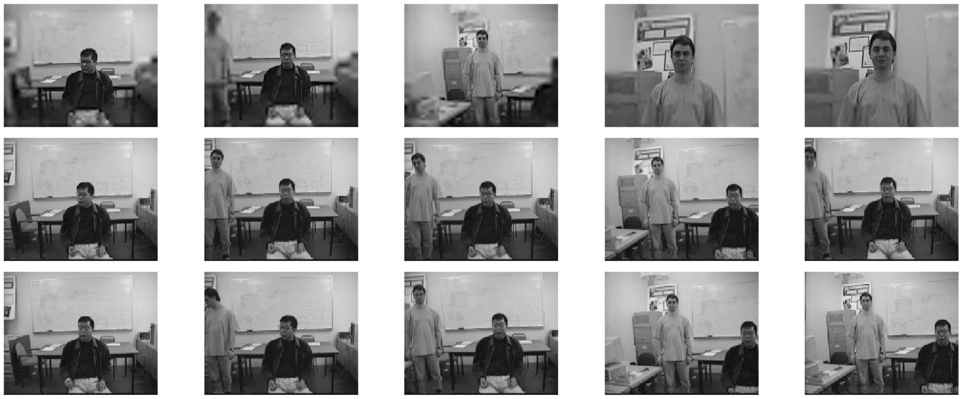


Fig. 3. Videos shot by the computer controlled camera in the same scenario with three different methods. Row 1: FPMA-FWT. Row 2: PMA-ORIGINAL. Row 3: PMA-WAVELET.

labeling processes are also conducted, and the criteria for camera pan/tilt and zooms are the same as those of the FPMA-FWT.

For the FPMA-FWT, only if a drastic motion occurs outside OFA can the saccade be issued. Conversely, for the other two methods many more unnecessary saccades are issued because these two methods are sensitive to motions in any position within the view. For the same reason, the other two methods have a difficult time to hold the object in the view, while the FPMA-FWT performs well in holding the object in OFA in the form of minor pan/tilt and zoom due to its variable allocation of attention to the FA and PA. Consequently, a much more satisfactory tracking behavior similar to the AVS is resulted from the FPMA-FWT.

To better illustrate the strength of the FPMA-FWT, a scenario is presented below as a comparison among these three different motion control schemes. The results are used to measure its success in simulating eye movements of the AVS. In this scenario, first only one man is in the view, then a second man is present and talks. Indeed the presented scenario, including gaze change and stabilization, is a typical scenario commonly encountered by any active systems. The responses of all three algorithms are illustrated in Fig. 3 and explained below.

The image sequences shot by the FPMA-FWT scheme in the given scenario is illustrated in the first row of Fig. 3. In the presence of the second man (Column 2), it responds in the right way—a saccade is issued and the newly appearing man is located on the central part of the view (Column 3). Next, due to the fact that the second man who has some head and shoulder motion stands far away from the camera, a zoom-in is issued by the FPMA-FWT since the moving area is around the OFA and qualifies the condition of the FPMA-FWT for zoom-in (Column 4), which is in agreement with the response of the AVS. Finally, because the second man is talking, which indicates only minor facial motions are involved, the holding motion of the camera should apply. Here the FPMA-FWT behaves as we desired: It locates the man on the center of each frame with only some minor pan/tilt motions (Column 5). This saccade–zoom–hold scenario emulates the response of the AVS well. Although, only one episode is presented in this paper, the active camera control has been autonomously and reliably working in our lab.

The image sequences shot by PMA-ORIGINAL and PMA-WAVELET are shown in the second and third row of Fig. 3 respectively. When the second man presents, both methods fail to issue a correct saccade since they are sensitive to the minor motion of the body of the first man, which contributes a large PMA, hence only a minor pan/tilt motion is issued to hold the first man. In the ensuing images, as results of the competition for the attention of the camera between the two men, some jerky saccades are provoked by these two schemes since they are alert to motions in any position in the view, no matter how small its magnitude is.

4 Recovering Camera Movements from FWT-Based Video

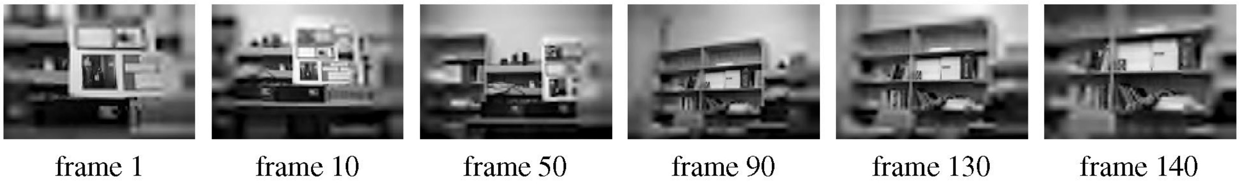
We shall demonstrate that pan/tilt/zoom camera movements can be effectively recovered from FWT-based videos under the assumption that the translating motion of the camera and object motion are too small compared with the pan/tilt/zoom velocity of the camera to be taken into account. Based on the formulas of the pin-hole camera image formation and the general 3D motion field, two properties can be derived for FWT-based representation.

1. *Zooming characteristic:* If the quotient q of the magnitude of the motion vector of any position over its distance from c is a constant, a camera zoom is detected. The zoom-out and zoom-in are determined by the occurrence of the Focus of Concentration (FOC) and Focus of Expansion (FOE).
2. *Pan/tilt characteristic:* The magnitudes of motion vectors $\vec{v} = (v_x, v_y)$ caused by camera pan/tilt movements are not of constant values across the entire image. For a pinhole camera, the errors η_x and η_y from a constant motion vector throughout the frame are of the value $|v_x| \sin^2 \varphi_x$ and $|v_y| \sin^2 \varphi_y$, respectively, where φ_x and φ_y are the corresponding viewing angles projected along the X and Y axes. An FWT-based approach reduces the impact of these errors greatly, thus, resulting in nearly constant motion vectors in the FWT-based frame.

To recover the motion field, a two-pass algorithm (TPA) based on Mean Field Theory under the Markov Random Field framework was proposed in [21], where two MRF's are used to model the motion vectors and unpredictable blocks. The configuration of motion vectors minimizing the energy functions is obtained as the detection results. More details about this algorithm can be found in [21]. The corresponding TPA based on the FWT representation, denoted by FWT-TPA, is then as follows: 1) First, perform the pixel-based TPA in S_8. 2) Propagate and refine motion vectors in the OFA and PPA of lower subbands in the FWT representation, now for each iteration the energy function to be minimized is formulated by considering the frequency information of the sibling locations in the three subbands. The motion vector field for level i obtained by carrying out the FWT-TPA is denoted as V_i.

4.1 Pan/Tilt/Zoom Camera Movements Recovery

Motion fields V_8 and V_1 as detected by the FWT-TPA are the foundations upon which our camera motion recovery scheme is built. Since V_8 has coarse motion vectors for each block, it thus



frame 1 frame 10 frame 50 frame 90 frame 130 frame 140

Fig. 4. The FWT-reconstructed image sequence with known camera motions.

encompasses the global information. Whereas V_1 has those corresponding to the FA of the highest possible resolution. The combination of these two fields yields a reliable indication of the camera motion. The other two intermediate fields, i.e.,V_4, V_2, are not employed in the ensuing camera motion recovery. Because S_8 is a low-passed version of the original frame by a scale factor of 8, the errors η_x (or η_y) in most areas, as described by the pan/tilt characteristic, are far less than one and can thus be ignored. However, in those regions close to the boundary this error may cause problems but these regions are merely very small portion of the DPA and their impact is ignored effectively by our use of robust statistical method in obtaining the motion fields. Thus, velocities in the DPA are also roughly constant. As for the FA, the viewing angles inside it are extremely small due to the choice of the FWT, recall that φ_x (φ_y) has small magnitude for neighboring frames, which induce a v_x (or v_y) of small magnitude, thereby η_x (or η_y) can be ignored and, thus, motion vectors in V_8 are approximately of constant value. Consequently to detect the camera pan/tilt, one can check the constancy of the motion vectors in V_1 and V_8. Our camera motion recovering algorithm, denoted as FREC, goes through the following three cases to determine different camera motions.

4.1.1 Pan/Tilt Only

As previously discussed, when only pan/tilt camera motions are involved, the constancy should be exhibited in both V_1 and V_8. Hence, in order to detect the pan/tilt, in V_1 we count the number of different motion vectors. If one motion vector counts overwhelmingly more than any others, it is then accepted as the camera pan/tilt motion vector (p_1, t_1). The pan/tilt camera motion vector (p_8, t_8) in V_8 can be obtained in the same manner. If $|p_1 - p_8 * 8| \leq \tau$, where τ is a small number (six is selected in our implementation), then (p_1, t_1) is accepted as the pan/tilt motion. The pan/tilt angle can be further obtained if f is given.

4.1.2 Zoom Only

Two steps are employed to detect the zoom motion:

1. According to the zoom characteristic, a camera zoom can be recovered by checking the *"constancy"* of q for each block/pixel in both V_1 and V_8. It is commonplace that in the results from the vision algorithms, many outliers are present. An ensemble of powerful tools from robust statistics can be employed to discount the adverse impacts of these outliers. To determine the constancy of q from noisy results of V_1 and V_8, the order statistics is utilized. Here, motion vectors are first sorted, then those vectors in the highest and lowest 20 percent are trimmed out. Next, the corresponding mean μ and standard deviation σ, are computed. If σ is extremely small, the motion vectors are considered to be constant, which is μ, and goto 2). Otherwise, exit from the zoom detection.
2. To distinguish the two cases of zoom, a vote on the direction of each motion vector $\vec{v}$ on the position m is incurred on V_1 and V_8 independently:
   a. calculate the angle $\alpha \in [0, \pi]$ between $\vec{v}$ and the vector $\vec{mc}$.
   b. If $\alpha < \frac{\pi}{4}$, the vote value val_m for $\vec{v}$ on m is 1; if $\alpha > \frac{3\pi}{4}$, val_m is -1; otherwise the corresponding val_m is 0.
   c. Calculate $Vote = \sum_m val_m$, suppose the number of pixels/blocks under consideration is N. On both V_1 and V_8 if $Vote > \frac{3N}{4}$ for their own N, a zoom-out is declared. If $Vote >$ -$\frac{3N}{4}$, the zoom-in is claimed. Otherwise, no zoom is claimed therein.

4.1.3 The Mixture of Pan/Tilt and Zoom

As a superposition of zoom and pan/tilt, for each pixel/block, the motion vector v projected on the two planes X and Y v_x and v_y are written as $v_x = z_x + p$ and $v_y = z_y + t$, where z_x and z_y are the contribution $\vec{z}$ of a camera zoom along x and y axis, respectively. To recover p, t, and the fact it is a zoom-in or zoom-out, according to the zoom characteristic, the following steps are employed:

1. For two motion fields V_1 and V_8 respectively:
   a. Add all v_x's for each pixel/block together, according to the zoom characteristic, z_x's for all pixels cancel each other, the mean of this sum can be viewed as the motion p contributed by pan. Apparently, t can be obtained in the same manner.
   b. From each v_x and v_y subtract the p and t, then we do the zoom detection on the remaining motions in the same manner as the last case.
2. Check the alignment of the zoom conclusion and the pan/tilt numbers for V_1 and V_8. If no significant difference is spotted, the recovered camera motion is declared as that estimated in V_1.

In order for the preceding algorithm to work, motion vectors for most pixels in V_1 and V_8 have to be estimated. Sufficient textures are thus needed in S_8 and the FA of the original resolution. Given that the resolution of S_8 is already one-eighth of that of the original image, and it is not unreasonable to assume that in the fovea area sufficient texture is always present, most motion vectors in V_1 and V_8 are thus computed using the FWT-TPA algorithm. The use of robust statistical schemes alleviate the negative impact of regions which lack textures in our decision about pan/tilt/zoom movements. Specifically, because of the use of order statistics, for scenes whose regions of insufficient texture are less than 20 percent of the entire view are effectively removed. Even if they are slightly more than 20 percent, the way for our algorithm to arrive at our conclusion is also able to alleviate its impact.

4.2 Experimental Results

In this section, to measure the performance of the proposed scheme for camera motion recovery, we feed videos with known camera motions into the computer, then camera movements computed by the recovery program is compared to known ones, whereby an effective indication of the efficacy of the recovery algorithm is obtained. Here, two other recovering algorithms are also implemented to present comparisons. The first is based on frames of the original resolutions, while the other is based on the wavelet transform of the whole image. The motion detection

TABLE 1
Motions Recovered by the Three Different Algorithms

Methods	motion 1	motion 2	motion 3	motion 4
ORIG	FAIL	(2, 4)	FAIL	FAIL
FREC	zoom-out(0.93)	(2, 4)	(12, 15)	zoom-in(0.46) & (-4, -3)
WAV	zoom-out(0.69)	(2, 3)	(12, 12)	FAIL

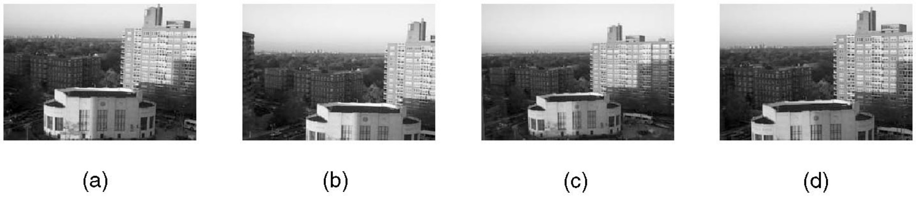


Fig. 5. Effects of pan-tilt-zoom camera movement recovery when there are significant depth differences. (a) and (b) Before and after a pan-tilt movement, (c) and (d) Before and after a zoom movement.

procedure used by both schemes are again TPA, denoted as ORIG and WAV, respectively.

As an illustration, six frames in a clip with known camera motions are presented in Fig. 4, which are reconstructed from their corresponding FWT representations. In this clip, the camera first underwent a zoom-out, then some pan/tilt motions, finally a zoom-in with pan/tilt was issued. The recovered motion with respect to four pairs of frames with known motions with the three different algorithms are listed in Table 1, where motion 1, 2, 3, and 4 corresponds to zoom-out ($q = 1.0$), pan/tilt (2, 4), pan/tilt (12, 15), and zoom-in ($q = 0.5$) with pan/tilt (-3, -4), respectively. Whereby, one can see that the FREC performs better than the other two methods. Here, one unit of pan/tilt position is 740 seconds arc ($\approx 0.2°$). In frame 50 and 90, it can be observed that regions of insufficient texture cover roughly 20 percent to 30 percent of the entire view, the proposed algorithm FREC still succeeds in arriving at the correct detection. Indeed, this is so achieved due to our use of robust statistical techniques in reaching the conclusion about pan/tilt/zoom movements. Therefore, our algorithm is reasonably resilient to regions of insufficient textures. Of course, in cases where more regions ($> \frac{1}{3}$) are of insufficient texture the robust statistical techniques adopted in our algorithm fails to discount their impact any more and, thus, unable to obtain the correct detection.

In order to further inspect the performance of the proposed algorithm, two pairs of outdoor frames with significant depth differences as shown in Fig. 5 are fed to our program to recover the camera movements. The pan/tilt camera movement from Fig. 5a) to Fig. 5b) is (2,3). The movement recovered by ORIG, WAV, and FREC is (1,2), (2,2), and (2,3). As can be seen that the viewing angles φ in the DPAs have posed negative impacts on the recovered movements of the ORIG and WAV; whereas, FREC is resilient to this problem based on the pan/tilt characteristic. For the second pair of frames, the camera movement from Fig. 5c) to Fig. 5d) is a zoom with $q = \frac{1}{3}$. The movement recovered by ORIG, WAV, and FREC is 0.19, 0.23, and 0.29, respectively. As previously reasoned the negative effects of those regions of insufficient texture are reduced due to the usage of the robust statistical techniques in our algorithm. More experiments along this line provide similar results as presented here. They further confirm the two characteristics and the fact that our algorithm is resilient to significant depth differences in the presence of camera pan/tilt/zoom motions.

Since the ORIG works on the frame of the original resolution, it can only work for extremely small pan/tilt and zoom because increasing the size of the search window is not only time consuming but also error-prone. Therefore, no more experimental results will be presented for it. More tests are conducted on the WAV and FREC schemes, both refining the motion vectors hierarchically. In Table 2, the overall results for pan/tilt, zoom, and a mixture of them are reported, each set of them contains 50 different frame pairs. Two measures ϵ_1 and ϵ_2 are defined to quantitatively compare performances of different methods. The accumulated error ϵ_1 for pan/tilt is defined as $\sum_i \frac{|v^i_{rh} - v^i_{kh}| + |v^i_{rv} - v^i_{kv}|}{2}$, where i ranges all frame pairs whose known camera movements are pan/tilt, and (v^i_{kh}, v^i_{kv}) is the known camera pan/tilt movement for frame pair i, and the corresponding recovered movement is (v^i_{rh}, v^i_{rv}). ϵ_2 defines the accumulated error for zooming movements $\epsilon_2 = \sum_j c_j$, where c_j is $\frac{q^j_r - q^j_k}{q^j_k}$ if a zoom is detected and one otherwise. j ranges over all frame pairs whose known camera movements are zooms, q^j_r and q^j_k are the recovered and known zooming quotients as defined in the preceding section.

For the mixture case, the corresponding accumulated error $\epsilon = \epsilon_1 + \epsilon_2$. Evidently for all three cases a smaller accumulated error indicates a better performance. From these tables it can be observed that the FREC method, as the least time-consuming one, performs consistently better than the other two recovering schemes, which further delineates the efficacy of the FWT as the frame representation.

5 Conclusion

In this paper, the Foveate Wavelet Transform (FWT) is proposed as a new technique to represent images with spatially variable resolutions (VR), which provides a desirable emulation of the animate visual systems. Compared with existing VR techniques, the advantages of the FWT are its linear-feature preservation, orientation selectivity and flexibility, although a hardware realization of the FWT has yet to be studied. As the first application of the FWT, an FPMA-FWT algorithm is proposed which can give rise to an efficient and effective active camera control scheme. Preliminary experimental results with encouraging performances have

TABLE 2
Complete Results of the Two Different Algorithms

Methods	Pan/tilt	zooming	combination
	ϵ_1	ϵ_2	$\epsilon_1 + \epsilon_2$
FREC	16.8	7.3	24.1
WAV	37.1	19.7	56.8

been demonstrated. It is also demonstrated that the FWT can be used to tackle the reverse problem of the active camera control, which is the camera pan/tilt/zoom motion recovery from raw videos, wherein the camera movements can be recovered in an effective and efficient manner. Information with regard to camera movements is of primary importance in the ensuing object segmentation and description, which further demonstrated the efficacy of the FWT.

In [19], the utility of the FWT in stereo active vision is examined, where depth information is added in determining the gaze control with encouraging performance. In the future, more refined algorithms will be developed to cope with more complex cases, e.g., relaxing the one-object constraint which may cause some problems in the interactions with the real world. More work needs to be done to deal with issues when vergent camera motion is also involved in order to explore a scene in a manner similar to that of [8]. Furthermore, it is also of great interest to devise FWT-based methods to process input visual data continuously in order to build a model of the scene under the framework of *vision as process* as developed by Crowley and Christensen [4]. In the upcoming video standard MPEG-7, the interactive manipulation and flexible representation of videos in terms of objects have a central role to play. The proposed FWT, as a representation which can facilitate practical active video acquisition and efficient video content descriptive schemes, can be of more utility as a candidate for practical real-time video representation in digital libraries.

REFERENCES

[1] C. Bandera and P. Scott, "Foveal Machine Vision Systems," *Proc. IEEE Int'l Conf. System, Man, and Cybernetics,* pp. 596-599, 1989.

[2] R.H.S. Carpenter, *Movements of the Eyes.* London: Pion, 1977.

[3] E.C. Chang and C. Yap, "A Wavelet Approach to Foveating Images," *Proc. ACM Symp. Computational Geometry,* pp. 397-399, 1997.

[4] J. Crowley and H.I. Christensen, *Vision as Process.* Berlin: Springer-Verlag, 1995.

[5] T. Darrell, G. Gordon, M. Harville, and J. Woodfill, "Integrated Person Tracking Using Stereo, Color, and Pattern Detection," *Proc. Computer Vision and Pattern Recognition (CVPR '98),* pp. 601-608, 1998.

[6] F. Ferrari, J. Nielsen, P. Questa, and G. Sandini, "Space Variant Imaging," *Sensor Rev.,* vol. 15, no. 2, pp. 17-20, 1995.

[7] M. Irani, R. Benny, and S. Peleg, "Computing Occluding and Transparent Motions," *Int'l J. Computer Vision,* vol. 12, no. 1, pp. 5-16, 1994.

[8] W.N. Klarquist and A.C. Bovik, "Fovea a Foveated Vergent Active Stereo Vision System for Dynamic Three-Dimensional Scene Recovery," *Trans. Robotics and Automation,* vol. 14, no. 5, pp. 755-780, 1998.

[9] S. Mallat and S. Zhong, "Characterization of Signals from Multiscale Edges," *Trans. Pattern Analysis and Machine Intelligence,* vol. 14, no. 7, pp. 710-732, July 1992.

[10] I.D Reid and D.W. Murray, "Active Tracking of Foveated Feature Clusters Using Affine Structure," *Int'l J. Computer Vision,* vol. 18, no. 1, pp. 1-20, 1996.

[11] G. Sandini and P. Dario, "Active Vision Based on Space-Variant Sensing," *Proc. Int'l Symp. Robotics Research,* pp. 75-83, 1990.

[12] G. Sandini et al. "Image-Based Personal Communication Using an Innovative Space-Variant CMOS Sensor," *Proc. IEEE Int'l Workshop on Robot and Human Comm.,* pp. 158-163, 1996.

[13] E.L. Schwartz, "Computational Anatomy and Functional Architecture of Striate Cortex: A Spatial Mapping Approach to Perceptual Coding," *Vision Research,* vol. 30, pp. 645-669, 1980.

[14] S.M. Smith and J.M. Brady, "A Scene Segmenter: Visual Tracking of Moving Vehicles," *Eng. Applications of Artificial Intelligence,* vol. 7, no. 2, pp. 191-204, 1994.

[15] M.J. Swain and M.A. Stricker, "Promising Directions in Active Vision," *Int'l J. Computer Vision,* vol. 11, no. 2, pp. 109-126, 1993.

[16] F. Tong and Z.N. Li, "Reciprocal-Wedge Transform for Space-Variant Sensing," *Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 6, pp. 500-511, 1995.

[17] J.Y.A. Wang and E.H. Adelson, "Representing Moving Images with Layers," *Trans. Image Processing,* vol. 3, no. 5, pp. 625-638, 1994.

[18] J. Wei, "Foveate Wavelet Transform and Its Applications in Digital Video Processing, Acquisition, and Indexing," PhD thesis, Simon Fraser Univ., 1998.

[19] J. Wei and Z.N. Li, "Efficient Disparity-Based Gaze Control with Foveate Wavelet Transform," *Proc. Int'l Conf. Intelligent Robots and Systems, (IROS '98).* pp. 866-871, 1998.

[20] J. Wei and Z.N. Li, "Foveate Wavelet Transform for Camera Motion Recovery from Videos," *Proc. Int'l Conf. Pattern Recognition, (ICPR,'98),* pp. 1445-1448, 1998.

[21] J. Wei and Z.N. Li, "The MAP-MRF Estimation of Motion Vectors Based on Mean Field Theory," *Trans. Circ. and Systems on Video Technology,* vol. 9, no. 6, pp. 960-972, 1999.

[22] C.F.R. Weiman and G. Chaikin, "Logarithmic Spiral Grids for Image Processing and Display," *Computer Graphics and Image Processing,* vol. 11, pp. 197-226, 1979.

[23] K. Wiebe, "Variable Resolution Vision: Biologically Motivated Foveal Compression and Prioritization," PhD thesis, Univ. of Alberta, 1996.

[24] K.J. Wiebe and A. Basu, "Modelling Ecologically Specialized Biological Visual Systems," *Pattern Recognition,* vol. 30, no. 10, pp. 1687-1703, 1997.

[25] J. Zhang and G.G. Hanauer, "The Application of Mean Field Theory to Image Motion Estimation," *Trans. Image Processing,* vol. 4, no. 1, pp. 19-32, 1995.