

# Global motion compensated visual attention-based video watermarking

Matthew Oakes,<sup>a</sup> Deepayan Bhowmik,<sup>b,\*</sup> and Charith Abhayaratne<sup>c</sup>

<sup>a</sup>University of Buckingham, Department of Applied Computing, Buckingham MK18 1EG, United Kingdom

<sup>b</sup>Sheffield Hallam University, Department of Computing, Sheffield S1 1WB, United Kingdom.

<sup>c</sup>University of Sheffield, Department of Electronic and Electrical Engineering, Sheffield S1 3JD, United Kingdom

**Abstract.** Imperceptibility and robustness are two key but complementary requirements of any watermarking algorithm. Low-strength watermarking yields high imperceptibility but exhibits poor robustness. High-strength watermarking schemes achieve good robustness but often suffer from embedding distortions resulting in poor visual quality in host media. This paper proposes a unique video watermarking algorithm that offers a fine balance between imperceptibility and robustness using motion compensated wavelet-based visual attention model (VAM). The proposed VAM includes spatial cues for visual saliency as well as temporal cues. The spatial modeling uses the spatial wavelet coefficients while the temporal modeling accounts for both local and global motion to arrive at the spatiotemporal VAM for video. The model is then used to develop a video watermarking algorithm, where a two-level watermarking weighting parameter map is generated from the VAM saliency maps using the saliency model and data are embedded into the host image according to the visual attentiveness of each region. By avoiding higher strength watermarking in the visually attentive region, the resulting watermarked video achieves high perceived visual quality while preserving high robustness. The proposed VAM outperforms the state-of-the-art video visual attention methods in joint saliency detection and low computational complexity performance. For the same embedding distortion, the proposed visual attention-based watermarking achieves up to 39% (nonblind) and 22% (blind) improvement in robustness against H.264/AVC compression, compared to existing watermarking methodology that does not use the VAM. The proposed visual attention-based video watermarking results in visual quality similar to that of low-strength watermarking and a robustness similar to those of high-strength watermarking. © The Authors. Published by SPIE under a Creative Commons Attribution 3.0 Unported License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI. [DOI: [10.1117/1.JEI.25.6.061624](https://doi.org/10.1117/1.JEI.25.6.061624)]

Keywords: video visual attention model; motion compensation; video watermarking; robustness; subjective visual quality evaluation.  
Paper 16380SS received May 1, 2016; accepted for publication Nov. 29, 2016; published online Dec. 20, 2016.

## 1 Introduction

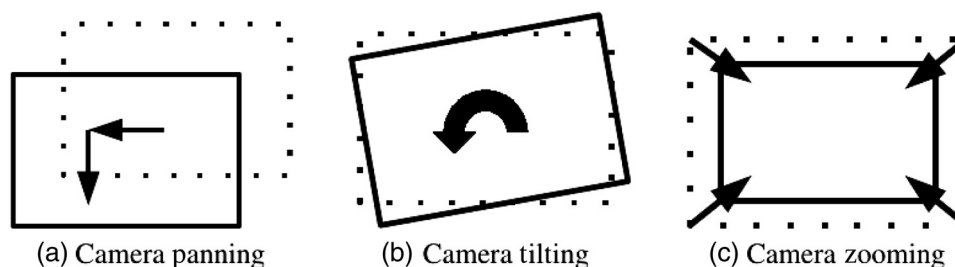
With the recent rapid growth of digital technologies, content protection now plays a major role within content management systems. Of the current systems, digital watermarking provides a robust and maintainable solution to enhance media security. The visual quality of the host media (often known as imperceptibility) and robustness are widely considered as the two main properties vital for a good digital watermarking system. They are complimentary to each other, hence it is challenging to attain the right balance between them. This paper proposes a new approach to achieve high robustness in watermarking while not affecting the perceived visual quality of the host media by exploiting the concepts of visual attention (VA).

The human visual system (HVS) is sensitive to many features which lead to attention being drawn toward specific regions in a scene and is a well-studied topic in psychology and biology.<sup>1,2</sup> VA is an important and complex biological process that helps to identify potential danger, e.g., prey, predators quickly in a cluttered visual world<sup>3</sup> as attention to one target leaves other targets less available.<sup>4</sup> Recently, a considerable effort was noticed in the literature in modeling

VA<sup>3</sup> that has applications in many related domains including media quality evaluation. Areas of visual interest stimulate neural nerve cells, causing the human gaze to fixate toward a particular scene area. The visual attention model (VAM) highlights these visually sensitive regions, which stimulates a neural response within the primary visual cortex.<sup>5</sup> Whether that neural vitalization be from contrast in intensity, a distinctive face, unorthodox motion, or a dominant color, these stimulative regions diverge human attention providing highly useful saliency maps within the media processing domain.

Human vision behavioral studies<sup>6</sup> and feature integration theory<sup>1</sup> have prioritized the combination of three visually stimulating low level features: intensity, color, and orientation which comprise the concrete foundations for numerous image domain saliency models.<sup>3,7,8</sup> Most saliency models often use multiresolution analysis.<sup>9–11</sup> Temporal features must be considered as moving objects and are more eye-catching than most static locations.<sup>12</sup> Work has seldom been directed toward video saliency estimation, in comparison to the image domain counterpart, as temporal feature consideration dramatically increases the overall VA framework complexity. Most typical video saliency estimation methodologies<sup>3,13–18</sup> exist as a supplementary extension from their image domain algorithms. Research estimating VA within video can also be derived from exploiting spatiotemporal cues,<sup>19,20</sup> structural tensors,<sup>21</sup> and optical flow.<sup>22</sup>

\*Address all correspondence to: Deepayan Bhowmik, E-mail: [deepayan.bhowmik@shu.ac.uk](mailto:deepayan.bhowmik@shu.ac.uk)



**Fig. 1** The three causes of global motion: camera panning, tilting, and zooming.

However, none of these algorithms explicitly captures the spatiotemporal cues that consider object motion between frames as well as the motion caused by camera movements. Motion within a video sequence can come from two categories namely, local motion and global motion. Local motion is the result of object movement within frames, which comprises all salient temporal data. One major feature associated with local motion is independence, so no single transformation can capture all local movement for the entire frame. Local motion can only be captured from successive frames differences if the camera remains motionless. On the contrary, global motion describes all motion in a scene based on a single affine transform from the previous frame and usually is a result of camera movement during a scene. The transform consists of three components, i.e., camera panning, tilting, and zooming or in image processing terms translation, rotation, and scaling. Figure 1 shows three causes for global motion. This paper proposes a new video VAM that accounts for local and global motions using a wavelet-based motion compensated temporal filtering framework. Compensating for any perceived camera movement reduces the overall effect of global movement so salient local object motion can be captured during scenes involving dynamic camera action.

A region of interest (ROI) dictates the most important visible aspects within media, so distortion within these areas will be highly noticeable to any viewer. The VAM computes such regions. This paper proposes a unique video watermarking algorithm exploiting the new video VAM. In frequency domain watermarking, the robustness of the watermarking is usually achieved by increasing the embedding strength. However, this results in visual distortions in the host media, thus a low imperceptibility of embedding. In the proposed method in this work, high watermark robustness without compromising the visual quality of the host media is achieved by embedding greater watermark strength within the less visually attentive regions within the media, as identified by the video VAM (in Sec. 2).

Related work includes defining an ROI<sup>23–28</sup> and increasing the watermark strength in the ROI to address cropping attacks. However, in these works, the ROI extraction was only based on foreground-background models rather than VAM. There are major drawbacks of such solutions: (a) increasing the watermark strength within eye catching frame regions is perceptually unpleasant as human attention will naturally be drawn toward any additional embedding artifacts, and (b) scenes exhibiting sparse salience will potentially contain extensively fragile or no watermark data. Sur et al.<sup>29</sup> proposed a pixel domain algorithm to improve embedding distortion using an existing visual saliency

model described in Ref. 3. However, the algorithm only discusses its limited observation on perceptual quality without considering any robustness.

Our previous work<sup>30,31</sup> shows the exploitation of image saliency in achieving image watermarking robustness. It is infeasible to simply extend the VA-based image domain algorithm into a frame-by-frame video watermarking scheme, as temporal factors must first be considered within the video watermarking framework. A viewer has unlimited time to absorb all information within an image, so potentially could view all conspicuous and visually uninteresting aspects in a scene. However, in a video sequence, the visual cortex has very limited processing time to analyze each individual frame. Human attention will naturally be drawn toward temporally active visually attentive regions. Thus the proposed motion compensated VAM is a suitable choice for VA-based video watermarking. By employing VA concepts within the digital watermarking, an increased overall robustness against adversary attacks can be achieved, while subjectively limiting any perceived visual distortions by the human eye. The concept of VA-based image and video watermarking was first introduced in our early work.<sup>30,32</sup> Recent work following this concept can be found in watermarking H.264 video<sup>33</sup> and application on cryptography.<sup>34</sup> On the contrary, in this paper, we propose a video watermark embedding strategy based on VA modeling that uses the same spatiotemporal decomposition used in the video watermarking scheme. In addition, the VAM compensates global motion in order to capture local motion into the saliency model.

Performances of our saliency model and the watermarking algorithms are separately evaluated by comparisons with existing schemes. Subjective tests for media quality assessment recommended by the International Telecommunication Union (ITU),<sup>35</sup> largely missing in the watermarking literature, are also conducted to complement the objective measurements. Major contributions of this paper are:

- A new motion compensated spatiotemporal video VAM that considers object motion between frames as well as global motions due to camera movement.
- New blind and nonblind video watermarking algorithms that are highly imperceptible and robust against compression attacks.
- Subjective tests that evaluate visual quality of the proposed watermarking algorithms.

The saliency model and the watermarking algorithms are evaluated using the existing video datasets described in Sec. 4.1. The initial concept of the motion compensated video attention model was reported earlier in the form of

a conference publication<sup>36</sup> while this paper discusses the proposed scheme in detail with an exhaustive evaluation and proposes a case study describing a new video watermarking scheme that uses the attention model.

## 2 Motion Compensated Video Visual Attention Model

The most attentive regions within media can be captured by exploiting and imposing characteristics from within the HVS. In this section, a method is proposed to detect any saliency information within a video. The proposed methods incorporate motion compensated spatiotemporal wavelet decomposition combined with HVS modeling to capture any saliency information. A unique approach combining salient temporal, intensity, color, and orientation contrasts formulate the essential video saliency methodology.

Physiological and psychophysical evidence demonstrates that visually stimulating regions occur at different scales within media<sup>37</sup> and the object motion within the scene.<sup>12</sup> Consequently, models proposed in this work exploit the identifiable multiresolution property of the wavelet transform that incorporates a motion compensation algorithm to generate the model. By exploiting the multiresolution spatiotemporal representation of the wavelet transform, VA is estimated directly from within the wavelet domain. The video saliency model is divided into three subsections. First, Sec. 2.1 describes the global motion compensation following the description of the spatial saliency model in Secs. 2.2 and 2.3 that illustrates the temporal saliency feature map generation. Finally, Sec. 2.4 combines the spatiotemporal model to estimate video visual saliency. An overall functional block diagram of our proposed model is shown in Fig. 2. For the spatial saliency model in this work, we adopted our image VAM proposed in Refs. 30 and 31.

### 2.1 Global Motion Compensated Frame Difference

Compensation for global motion is dependent upon homogeneous motion vector (MV) detection, consistent throughout the frame. Figure 3 considers the motion estimation between two consecutive frames, taken from the coastguard sequence. A fixed block size based on the frame resolution determines the number of MV blocks. The magnitude and phase of the MVs are represented by the size and direction of the arrows, respectively, whereas the absence of an arrow portrays an MV of zero. First, it is assumed there is a greater percentage of pixels within moving objects than in the background, so large densities of comparative MVs are the result of dynamic camera action. To compensate for camera panning, the entire reference frame is spatially translated by the most frequent MV, the global camera MV,  $\vec{M}_{\text{global}}$ . This



Fig. 3 Motion block estimation.

process is applied prior to the 2-D +  $t$  wavelet decomposition to deduce global motion compensated saliency estimation. The global motion compensation is described in Eq. (1)

$$\vec{M}_{\text{object}} = \vec{M}_{\text{total}} - \vec{M}_{\text{global}}, \quad (1)$$

where  $\vec{M}_{\text{object}}$  is the local object MV and  $\vec{M}_{\text{total}}$  is the complete combined MV.

Compensating for other camera movement can be achieved by searching for a particular pattern of MVs. For example, a circular MV pattern will determine camera rotation and all MVs converging or diverging from a particular point will govern camera zooming. An iterative search over all possible MV patterns can cover each type of global camera action.<sup>38</sup> Speeded up robust features detection<sup>39</sup> could be used to directly align key feature points between consecutive frames, but this would be very computationally exhaustive. This model only requires a fast rough global motion estimate to neglect the effect of global camera motion on the overall saliency map.

### 2.2 Spatial Saliency Model

As the starting point in generating the saliency map from a color image/frame, RGB color space is converted to YUV color spectral space as the latter exhibits prominent intensity variations through its luminance channel Y. First, the two-dimensional (2-D) forward discrete wavelet transform (FDWT) is applied on each Y, U, and V channel to decompose them in multiple levels. The 2-D FDWT decomposes an image in frequency domain expressing coarse grain approximation of the original signal along with three fine grain orientated edge information at multiple resolutions. Discrete wavelet transform (DWT) captures horizontal, vertical, and diagonal contrasts within an image portraying prominent edges in various orientations. Due to the dyadic nature of the multiresolution wavelet transform, the image resolutions are decreased after each wavelet decomposition iteration. This is useful in capturing both short and long structural information at different scales and useful for saliency computation. The absolute values of the wavelet coefficients are normalized so that the overall saliency contributions come

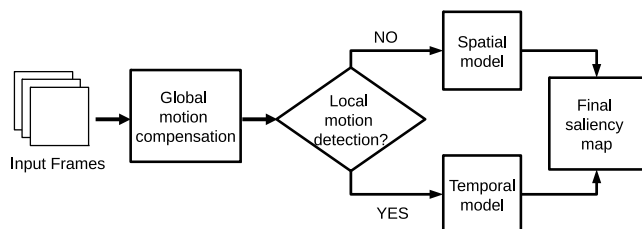


Fig. 2 Proposed video VAM functional block diagram.



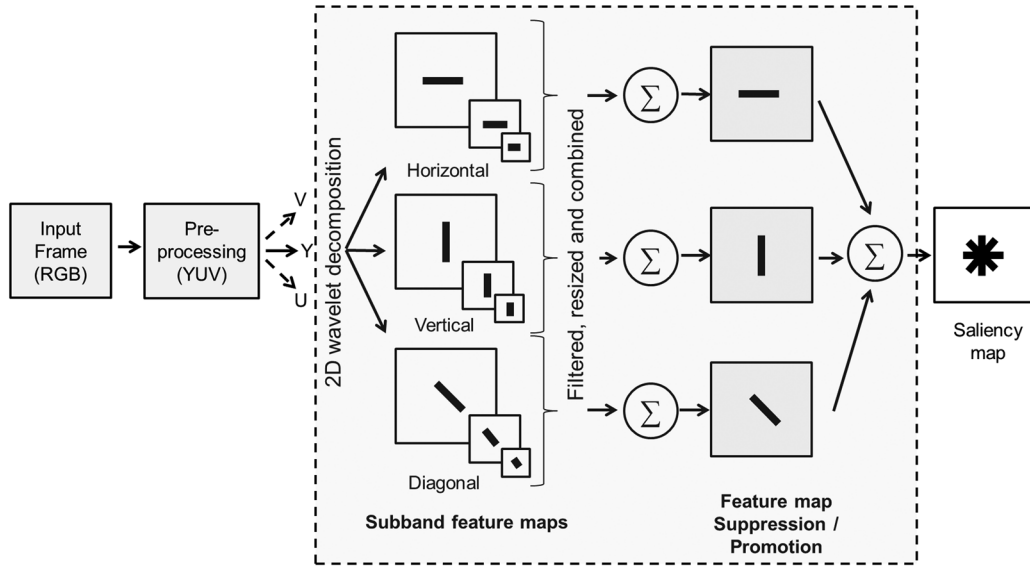


Fig. 4 Overall functional diagram of the spatial visual saliency model.

from each subband and prevent biasing toward the finer scale subbands. An average filter is also applied to remove unnecessary finer details. To provide full resolution output maps, each of the high frequency subbands is consequently interpolated up to full frame resolution. The interpolated subband feature maps,  $LH_i$  (horizontal),  $HL_i$  (vertical), and  $HH_i$  (diagonal),  $i \in \mathbb{N}_1$ , for all decomposition levels  $L$  are combined by a weighted linear summation as

$$\begin{aligned} LH_{1 \dots L_X} &= \sum_{i=1}^L LH_i * \tau_i, \\ HL_{1 \dots L_X} &= \sum_{i=1}^L HL_i * \tau_i, \\ HH_{1 \dots L_X} &= \sum_{i=1}^L HH_i * \tau_i, \end{aligned} \quad (2)$$

where  $\tau_i$  is the subband weighting parameter and  $LH_{1 \dots L_X}$ ,  $HL_{1 \dots L_X}$ , and  $HH_{1 \dots L_X}$  are the subband feature maps for a given spectral channel  $X$ , where  $X \in \{Y, U, V\}$ .

A feature map promotion and suppression steps follow next as shown in Eq. (3). If  $\bar{m}$  is the average of local maxima present within the feature map and  $M$  is the global maximum, the promotion and suppression normalization is achieved by

$$\begin{aligned} \overline{LH}_X &= LH_{1 \dots L_X} * (M - \bar{m})^2, \\ \overline{HL}_X &= HL_{1 \dots L_X} * (M - \bar{m})^2, \\ \overline{HH}_X &= HH_{1 \dots L_X} * (M - \bar{m})^2, \end{aligned} \quad (3)$$

where  $\overline{LH}_X$ ,  $\overline{HL}_X$ , and  $\overline{HH}_X$  are the normalized set of subband feature maps.

Finally, the overall saliency map,  $S$ , is generated by

$$S = \sum_{X \in \{Y, U, V\}} w_X * S_X, \quad (4)$$

where  $w_X$  is the weight given to each spectral component and  $S_X$  is the saliency map for each spectral channel ( $Y, U, V$ ), which is computed as follows:

$$S_X = \overline{LH}_X + \overline{HL}_X + \overline{HH}_X. \quad (5)$$

Finally, the overall map is generated by using a weight summation of all color channels as shown in Fig. 4.

## 2.3 Temporal Saliency Model

### 2.3.1 2-D + t wavelet domain

We extend our spatial saliency model toward video domain saliency logically by utilizing a three-dimensional wavelet transform. Video coding research provides evidence that differing texture and motion characteristics occur after wavelet decomposition from the  $t + 2$ -D domain<sup>40</sup> and incorporating its alternative technique, the 2-D +  $t$  transform.<sup>41,42</sup> The  $t + 2$ -D domain decomposition compacts most of the transform coefficient energy within the low frequency temporal subband and provides efficient compression within the temporal high frequency subbands. Vast quantities of the high frequency coefficients have zero magnitude, or very close, which is unnecessary for the transforms' usefulness within this framework. Alternatively, 2-D +  $t$  decomposition produces greater transform energy within the higher frequency components, i.e., a greater amount of larger and nonzero coefficients and reduces computational complexity to a great extent. A description of reduced computational complexity by using 2-D +  $t$  compared to  $t + 2$ -D can be found in Ref. 42. Therefore, in this work we have used a 2-D +  $t$  decomposition as shown in Fig. 5 (for three levels of spatial followed by one level of temporal Haar wavelet decomposition).

### 2.3.2 Temporal saliency feature map

To acquire accurate video saliency estimation, both spatial and temporal features within the wavelet transform are considered. The wavelet-based spatial saliency model, described in Sec. 2.2, constitutes the spatial element for the video

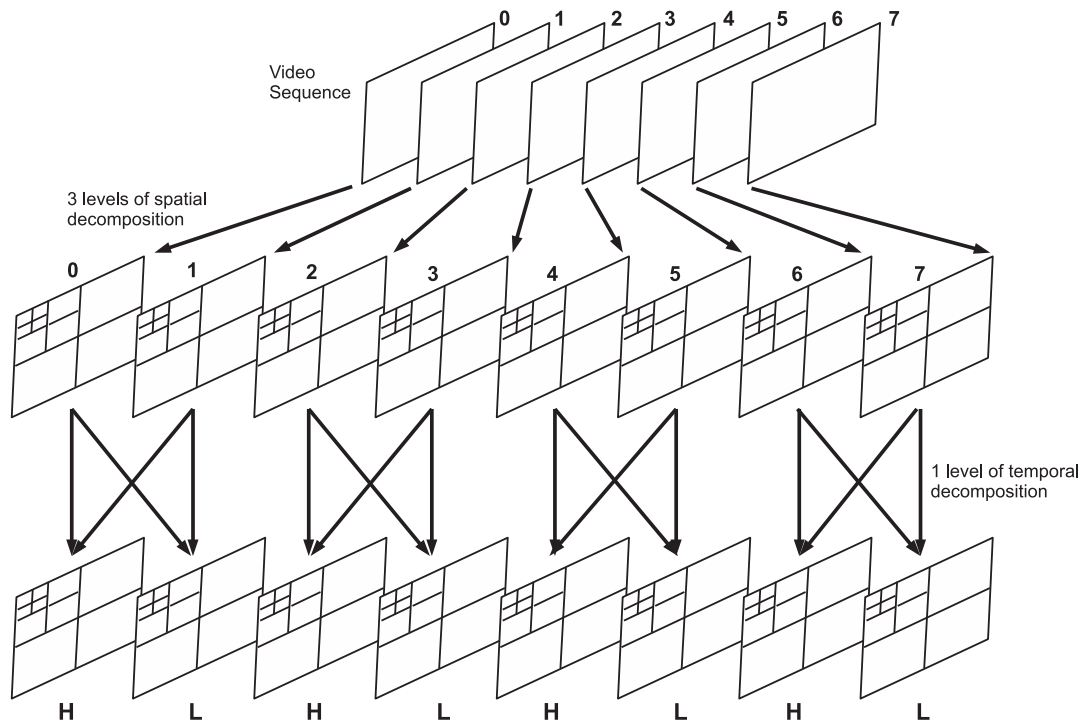


Fig. 5 2-D +  $t$  wavelet decomposition.

saliency model, whereas this section concentrates upon establishing temporal saliency maps,  $S_{\text{Temp}}$ .

Similar methodology to expose temporal conspicuousness is implemented in comparison to the spatial model in Sec. 2.2. First, the existence of any palpable local object motion is determined within the sequence. Figure 6 shows the histograms of two globally motion compensated frames. Global motion is any frame motion due to camera movement, whether that be panning, zooming, or rotation (see Sec. 2.1). Change within lighting, noise, and global motion compensation error account for the peaks present within Fig. 6(a), whereas the contribution from object movement is also present within Fig. 6(b). A local threshold,  $T$ , segments frames containing sufficiently noticeable local motion,  $\mathcal{M}$ , from an entire sequence. If  $F_1$  and  $F_2$  are consecutive 8-bit luma frames within the same sequence, Eq. (6) classifies temporal frame dynamics using frame difference  $D$

$$D(x, y) = |F_1(x, y) - F_2(x, y)|. \quad (6)$$

From the histograms shown within Figs. 6(a) and 6(b), a local threshold value of  $T = D_{\text{max}}/10$  determines motion classification, where  $D_{\text{max}}$  is the maximum possible frame pixel difference, and  $T$  is highlighted by a red dashed line within both figures. A 0.5 percent error ratio of coefficients representing local motion  $\mathcal{M}$  must be greater than  $T$  to reduce frame misclassification. For each temporally active frame, the Y channel renders sufficient information to estimate salient object movement without considering the U and V components.

The  $S_{\text{Temp}}$  methodology bears a distinct similarity to the spatial domain approach as the high pass temporal subbands:  $LHt_i$ ,  $HLt_i$ , and  $HHt_i$ , for  $i$  levels of spatial decomposition, combine after full 2-D +  $t$  wavelet decomposition, which is shown in Fig. 5. The decomposed data are forged using

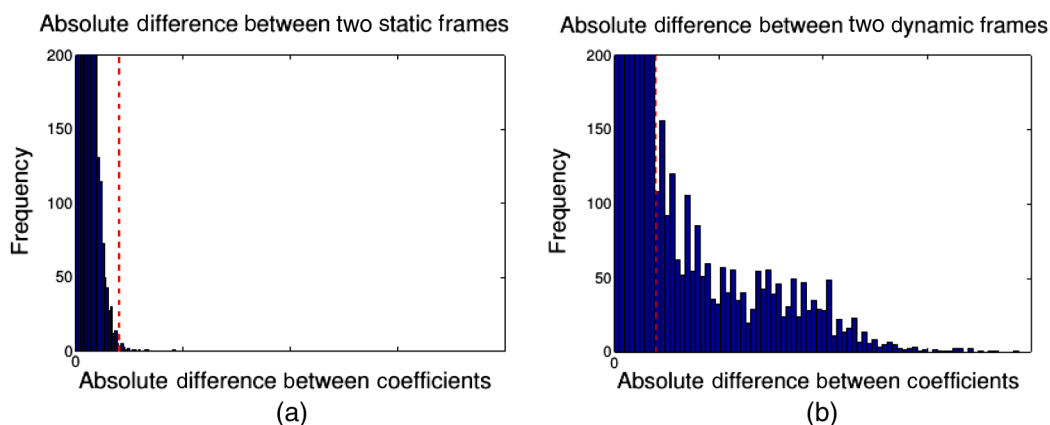


Fig. 6 Difference frames after global motion compensation: a sequence (a) without local motion and (b) containing local motion.

comparable logic as Eq. (2), as all transformed coefficients are segregated into 1 of 3 temporal subband feature maps. This process is described as

$$\begin{aligned} LHt_t &= \sum_{i=1}^n (|LHt_i|^{12^i} * \tau_i), \\ HLt_t &= \sum_{i=1}^n (|HLt_i|^{12^i} * \tau_i), \\ HHt_t &= \sum_{i=1}^n (|HHt_i|^{12^i} * \tau_i), \end{aligned} \quad (7)$$

where  $LHt_t$ ,  $HLt_t$ , and  $HHt_t$  are the temporal  $LH$ ,  $HL$ , and  $HH$  combined feature maps, respectively. The method captures any subtle conspicuous object motion in horizontal, vertical, and diagonal directions. This subsequently fuses the coefficients into a meaningful visual saliency approximation by merging the data across multiple scales.  $S_{Temp}$  is finally generated from

$$S_{Temp} = LHt_t + HLt_t + HHt_t. \quad (8)$$

## 2.4 Spatial-Temporal Saliency Map Combination

The spatial and temporal maps are combined to form an overall saliency map. The primary visual cortex is extremely sensitive to object movement so if enough local motion is detected within a frame, the overall saliency estimation is dominated by any temporal contribution with respect to local motion  $\mathcal{M}$ . Hence, the temporal weightage parameter,  $\gamma$ , determined from Eq. (6) is calculated as

$$\gamma = \begin{cases} 1 & \text{if } \mathcal{M} > T, \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

If significant motion is detected within a frame, the complete final saliency map comprises solely from the temporal feature. Previous studies support this theory, providing evidence that local motion is the most dominant feature within low level VA.<sup>43</sup> Consequently, if no local motion is detected with a frame, the spatial model contributes toward the final saliency map in its entirety, hence  $\gamma$  is a binary variable. The equation forging the overall saliency map is

$$S_{Final} = S_{Temp} * \gamma + S_{Spat} * (1 - \gamma), \quad (10)$$

where  $S_{Spat}$ ,  $S_{Temp}$ , and  $S_{Final}$  are the spatial, temporal, and combined overall saliency maps, respectively. An overall diagram for the entire proposed system is shown in Fig. 2.

## 3 Visual Attention-Based Video Watermarking

We propose an algorithm that provides a solution toward blind and nonblind VA-based video watermarking. The video saliency model described in Sec. 2 is utilized within the video watermarking framework to determine the watermarking embedding strength. Coinciding with the previous video VA model, watermark data are embedded within the 2-D +  $t$  wavelet domain as outlined in Sec. 2.3.1. The VAM identifies the ROI most perceptive to human vision, which is a highly exploitable property when designing

watermarking systems. The subjective effect of watermark embedding distortion can be greatly reduced if any artifacts occur within inattentive regions. By incorporating VA-based characteristics within the watermarking framework, algorithms can provide a retained media visual quality and increased overall watermark robustness, compared with the methodologies that do not exploit the VA. This section proposes two (blind and nonblind) new video watermarking approaches that incorporate the VAM. In both scenarios, a content-dependent saliency map is generated which is used to calculate the region adaptive watermarking strength parameter alpha,  $\alpha \in [0,1]$ . A lower and higher value of  $\alpha$  in salient regions and nonsalient regions, respectively, ensures higher imperceptibility of the watermarked image distortions while keeping greater robustness.

### 3.1 Watermarking Algorithms

At this point, we describe the classical wavelet-based watermarking schemes without considering the VAM and subsequently propose the new approach that incorporates the saliency model. Frequency-based watermarking, more precisely wavelet domain watermarking, methodologies are highly favored in the current research era. The wavelet domain is also compliant within many image coding, e.g., JPEG2000<sup>44</sup> and video coding, e.g., motion JPEG2000, motion-compensated embedded zeroblock coding (MC-EZBC),<sup>45</sup> schemes, leading to smooth adaptability within modern frameworks. Due to the multiresolution decomposition and the property to retain spatial synchronization, which are not provided by other transforms (the discrete cosine transform for example), the DWT provides an ideal choice for robust watermarking.<sup>46–55</sup>

The FDWT is applied on the host image before watermark data are embedded within the selected subband coefficients. The inverse discrete wavelet transform reconstructs the watermarked image. The extraction operation is performed after the FDWT. The extracted watermark data are compared to the original embedded data sequence before an authentication decision verifies the watermark presence. A wide variety of potential adversary attacks, including compression and filtering, can occur in an attempt to distort or remove any embedded watermark data. A detailed discussion of such watermarking schemes can be found in Ref. 56.

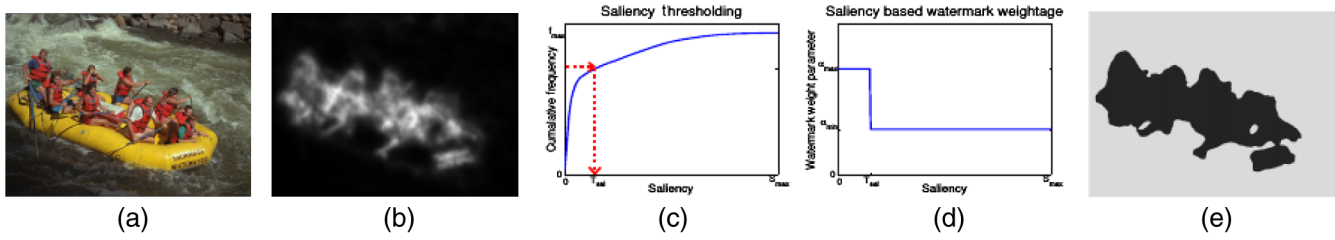
#### 3.1.1 Nonblind watermarking

Magnitude-based multiplicative watermarking<sup>34,51,53,57–59</sup> is a popular choice when using a nonblind watermarking system due to its simplicity. Wavelet coefficients are modified based on the watermark strength parameter,  $\alpha$ , the magnitude of the original coefficient,  $C(m, n)$ , and the watermark information,  $W(m, n)$ . The watermarked coefficients,  $C'(m, n)$ , are obtained as follows:

$$C'(m, n) = C(m, n) + \alpha W(m, n) C(m, n). \quad (11)$$

$W(m, n)$  is derived from a pseudorandom binary sequence,  $b$ , using weighting parameters,  $W_1$  and  $W_2$  (where  $W_2 > W_1$ ), which are assigned as follows:

$$W(m, n) = \begin{cases} W_2 & \text{if } b = 1 \\ W_1 & \text{if } b = 0. \end{cases} \quad (12)$$



**Fig. 7** (a) Example host image, (b) VAM saliency map (saliency is proportional to the gray scale), (c) cumulative saliency histogram, (d)  $\alpha$  step graph, and (e)  $\alpha$  strength map (dark corresponds to low strength).

To obtain the extracted watermark,  $W'(m, n)$ , Eq. (11) is rearranged as

$$W'(m, n) = \frac{C'(m, n) - C(m, n)}{\alpha C(m, n)}. \quad (13)$$

Since the nonwatermarked coefficients,  $C(m, n)$ , are needed for comparison, this results in nonblind extraction. A threshold limit of  $T_w = \frac{W_1 + W_2}{2}$  is used to determine the extracted binary watermark  $b'$  as follows:

$$b' = \begin{cases} 1 & \text{if } W'(m, n) \geq T_w \\ 0 & \text{if } W'(m, n) < T_w. \end{cases} \quad (14)$$

### 3.1.2 Blind watermarking

Quantization-based watermarking<sup>52,54–63</sup> is a blind scheme which relies on modifying various coefficients toward a specific quantization step. As proposed in Ref. 52, the algorithm is based on modifying the median coefficient toward the step size,  $\delta$ , by using a running nonoverlapping  $3 \times 1$  pixels window. The altered coefficient must retain the median value of the three coefficients within the window after the modification. The equation calculating  $\delta$  is described as follows:

$$\delta = \alpha \frac{(C_{\min}) + (C_{\max})}{2}, \quad (15)$$

where  $C_{\min}$  and  $C_{\max}$  are the minimum and maximum coefficients, respectively. The median coefficient,  $C_{\text{med}}$ , is quantized towards the nearest step, depending on the binary watermark,  $b$ . The extracted watermark,  $b'$ , for a given window position, is extracted by

$$b' = \left[ \frac{C_{\max} - C_{\text{med}}}{\delta} \right] \% 2, \quad (16)$$

where  $\%$  denotes the modulo operator to detect an odd or even number and  $C_{\text{med}}$  is the median coefficient value within the  $3 \times 1$  pixels window.

### 3.1.3 Authentication of extracted watermarks

Authentication is performed by comparison of the extracted watermark with the original watermark information and computing closeness between the two in a vector space. Common authentication methods are defined by calculating the similarity correlation or Hamming distance,  $H$ , between the original embedded and extracted watermark as follows:

$$H(b, b') = \frac{1}{N} \sum b \oplus b', \quad (17)$$

where  $N$  represents the length of the watermark sequence and  $\oplus$  is the XOR logical operation between the respective bits.

### 3.2 Saliency Map Segmentation

This section presents the threshold-based saliency map segmentation which is used for adapting the watermarking algorithms described in Sec. 3.1 in order to change the watermark strength according to the underlying VA properties. Figures 7(a) and 7(b) show an example original host frame and its corresponding saliency map, respectively, generated from the proposed methodology in Sec. 2. In Fig. 7(b), the light and dark regions within the saliency map represent the visually attentive and nonattentive areas, respectively. At this point, we employ thresholding to quantize the saliency map into coarse saliency levels as fine granular saliency levels are not important in the proposed application. In addition, that may also lead to reducing errors in saliency map regeneration during watermark extraction as follows. Recalling blind and nonblind watermarking schemes in Sec. 3.1, the host media source is only available within nonblind algorithms. However in blind algorithms, identical saliency reconstruction might not be possible within the watermark extraction process due to the coefficient values changed by watermark embedding as well as potential attacks. Thus, the saliency map is quantized using thresholds leading to regions of similar visual attentiveness. The employment of a threshold reduces saliency map reconstruction errors, which may occur as a result of any watermark embedding distortion, as justified further in Sec. 3.4.

The thresholding strategy relies upon a histogram analysis approach. Histogram analysis depicts automatic segmentation of the saliency map into two independent levels by employing the saliency threshold,  $T_s$ , where  $s \in S$  represents the saliency values in the saliency map,  $S$ . In order to segment highly conspicuous locations within a scene, first, the cumulative frequency function,  $f$ , of the ordered saliency values,  $s$ , (from 0 to the maximum saliency value,  $s_{\max}$ ) is considered. Then  $T_s$  is chosen as

$$T_s = f^{-1}(p * f_{\max}), \quad (18)$$

where  $p$  corresponds to the percentage of the pixels that can be set as the least attentive pixels and  $f_{\max} = f(s_{\max})$  corresponds to the cumulative frequency corresponding to the maximum saliency value,  $s_{\max}$ . An example of a cumulative



frequency plot of a saliency map and finding  $T_s$  for  $p = 0.75$  is shown in Fig. 7(c).

Saliency-based thresholding enables determining the coefficients' eligibility for a low- or high-strength watermarking. To ensure VA-based embedding, the watermark weighting parameter strength,  $\alpha$ , in Eqs. (11) and (15), is made variable  $\alpha(j, k)$ , dependent upon  $T_s$ , as follows:

$$\alpha(j, k) = \begin{cases} \alpha_{\max} & \text{if } s(j, k) < T_s, \\ \alpha_{\min} & \text{if } s(j, k) \geq T_s, \end{cases} \quad (19)$$

where  $\alpha(j, k)$  is the adaptive watermark strength map giving the  $\alpha$  value for the corresponding saliency at a given pixel coordinate  $(j, k)$ . The watermark weighting parameters,  $\alpha_{\min}$  and  $\alpha_{\max}$ , correspond to the high and low strength values, respectively, and their typical values are determined from the analysis within Sec. 3.3. As shown in Fig. 7(d), the most and the least salient regions are given watermark weighting parameters of  $\alpha_{\min}$  and  $\alpha_{\max}$ , respectively. An example of the final VA-based alpha watermarking strength map is shown in Fig. 7(e), where a brighter intensity represents an increase in  $\alpha$ .

### 3.3 Watermark Embedding Strength Calculation

The watermark weighting parameter strengths,  $\alpha_{\max}$  and  $\alpha_{\min}$ , can be calculated from the visible artifact peak signal-to-noise ratio (PSNR) limitations within the image. Visual distortion becomes gradually noticeable as the overall PSNR drops below 40 ~ 35 dB,<sup>64</sup> so minimum and maximum PSNR requirements are set to approximate 35 and 40 dB, respectively, for both the blind and nonblind watermarking schemes. These PSNR limits ensure a maximum amount of data can be embedded into any host image to enhance watermark robustness without substantially distorting the media quality. Therefore, it is sensible to incorporate PSNR in determining the watermark strength parameter  $\alpha$ .

Recall that PSNR, which measures the error between two images with dimensions  $X \times Y$ , is expressed on the pixel domain as follows:

$$\text{PSNR}(I, I') = 10 \log \left\{ \frac{M^2}{\frac{1}{XY} \sum_{j=1}^X \sum_{k=1}^Y [I'(j, k) - I(j, k)]^2} \right\}, \quad (20)$$

where  $M$  is the maximum coefficient value of the data, and  $I(j, k)$  and  $I'(j, k)$  are the original and watermarked image pixel values at  $(j, k)$  indices, respectively. Considering the use of orthogonal wavelet kernels and the Parseval's theorem, the mean square error in the wavelet domain due to watermarking is equal to the mean square error in the spatial domain.<sup>48</sup> Therefore, Eq. (20) can be redefined on the transform domain for nonblind magnitude-based multiplicative watermarking, shown in Eq. (11), as follows:

$$\text{PSNR}(I, I') = 10 \log \left\{ \frac{M^2}{\frac{1}{XY} \sum_{m=1}^X \sum_{n=1}^Y [\alpha W(m, n) C(m, n)]^2} \right\}. \quad (21)$$

By rearranging for  $\alpha$ , an expression determining the watermark weighting parameter depending on the desired PSNR

value is derived for nonblind watermarking in Eq. (22) as follows:

$$\alpha = \frac{M}{\sqrt{\frac{10^{(\text{PSNR}(I, I')/10)}}{XY} \sum_{m=1}^X \sum_{n=1}^Y [W(m, n) C(m, n)]^2}}. \quad (22)$$

Similarly, for the blind watermarking scheme described in Sec. 3.1.2, PSNR in the transform domain can be estimated by substituting the median and modified median coefficients,  $C_{(\text{med})}$  and  $C'_{(\text{med})}$ , respectively, in Eq. (20). Then subsequent rearranging results in an expression for the total error in median values, in terms of the desired PSNR as follows:

$$\sum_{m=1}^X \sum_{n=1}^Y (C'_{(\text{med})} - C_{(\text{med})})^2 = XY \frac{M^2}{10^{(\text{PSNR}/10)}}. \quad (23)$$

Equation (23) determines the total coefficient modification for a given PSNR requirement, hence it is used for  $\alpha$  in Eq. (15).

### 3.4 Saliency Map Reconstruction

For nonblind watermarking, the host data are available during watermark extraction so an identical saliency map can be generated. However, a blind watermarking scheme requires the saliency map to be reconstructed based upon the watermarked media, which may have gotten pixel values slightly different from the original host media. Thresholding the saliency map into two levels, as described in Sec. 3.2, ensures high accuracy within the saliency model reconstruction for blind watermarking. Further experimental objective analysis reveals that the use of thresholding improves the saliency coefficients match up to 99.4% compared to approximately only 55.6% of coefficients when thresholding was not used, hence reconstruction errors are greatly reduced.

## 4 Experimental Results and Discussion

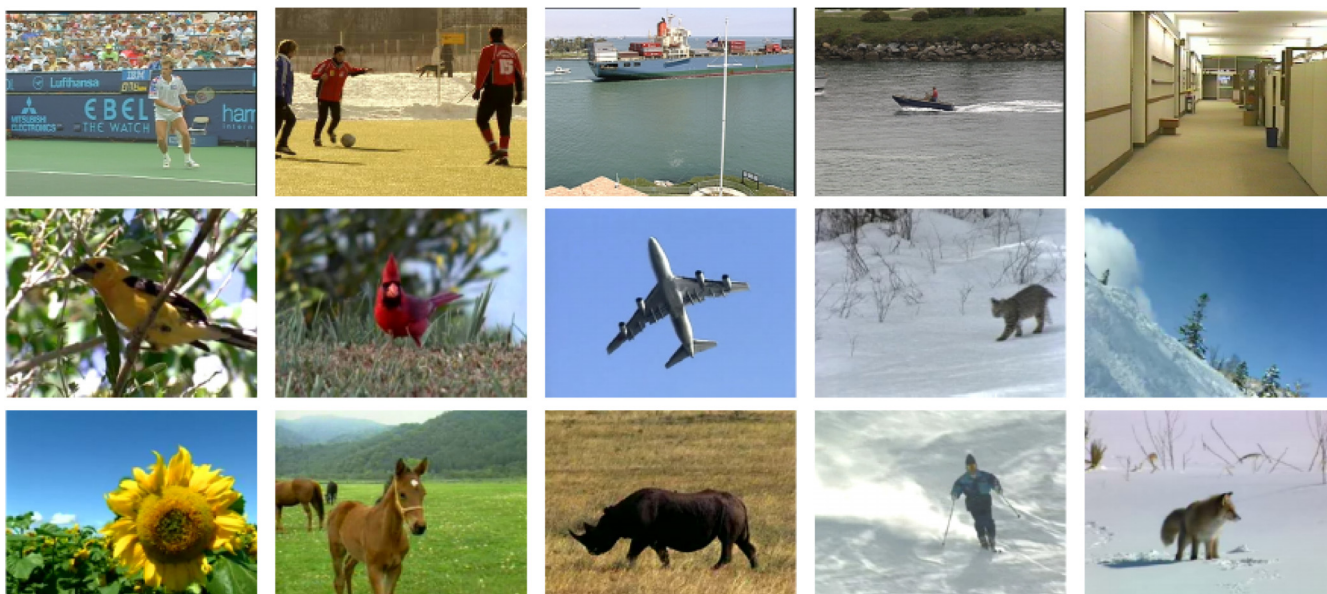
The performance of the proposed video VA method and its application in robust video watermarking is presented and discussed in this section. The video VAM is evaluated in terms of the accuracy with respect to the ground truth and computational time in Sec. 4.1. The video VA-based watermarking is evaluated in terms of embedding distortion and robustness to compression in Sec. 4.2.

### 4.1 Visual Attention Model Evaluation

For attention model evaluation, the video dataset is taken from the literature,<sup>65</sup> which is comprised of 15 video sequences, containing over 2000 frames in total. Ground truth video sequences have been generated from the database by subjective testing. A thumbnail from each of the 15 test sequences are shown in Fig. 8. Common test set parameters for VAM and later in watermarking, used throughout all performed experiments, include: the orthogonal Daubechies length 4 (D4) wavelet for three levels of 2-D spatial decomposition and one level of motion compensated temporal Haar decomposition.

Experimental results demonstrate the model performance against the existing state-of-the-art methodologies. The proposed algorithm is compared with the Itti,<sup>15</sup> dynamic,<sup>66</sup> and





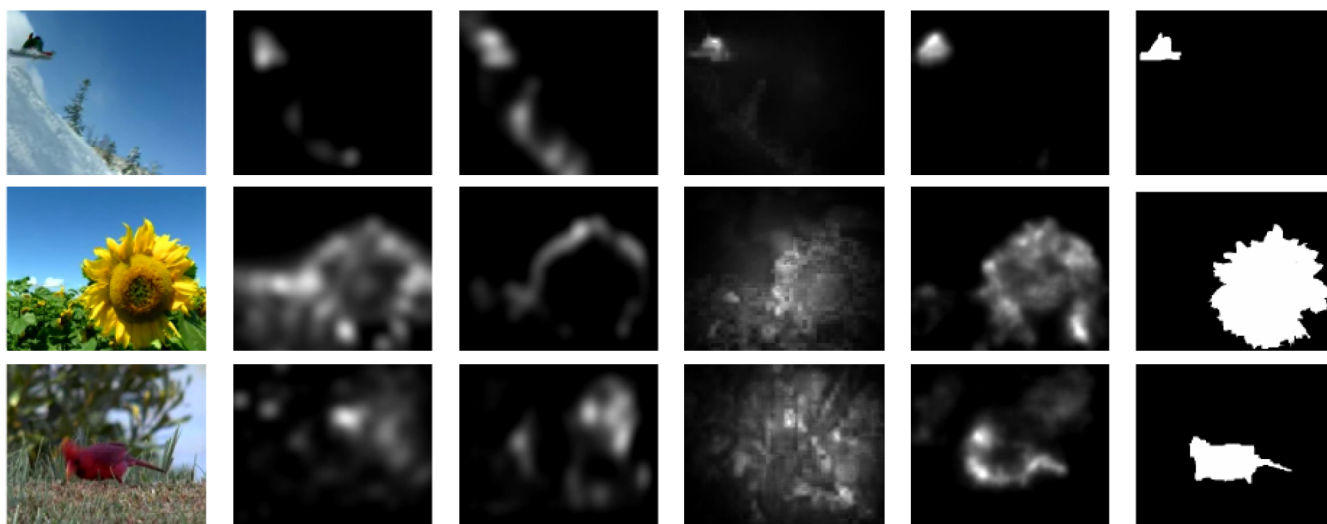
**Fig. 8** Video database: 15 thumbnails for each sequence.

Fang<sup>19</sup> video VAMs, in terms of accurate salient region detection and computational efficiency. The Itti framework is seen as the foundation and benchmark used for VA model comparison, whereas the dynamic algorithm is dependent upon locating energy peaks within incremental length coding. A more recent Fang algorithm uses a spatiotemporally adaptive entropy-based uncertainty weighting approach.

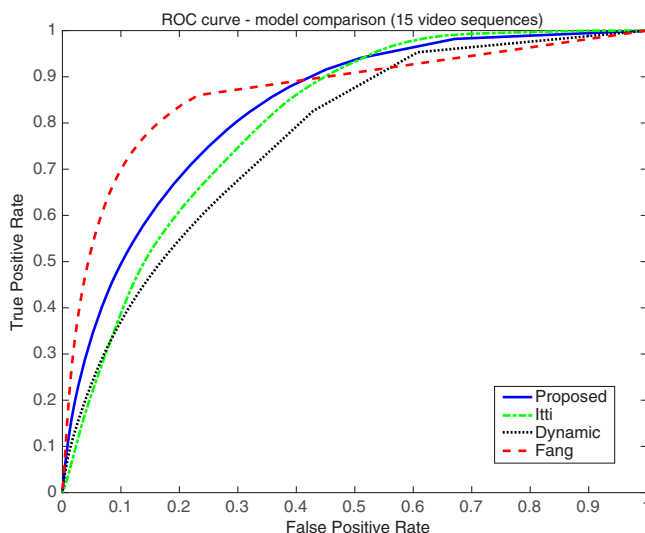
Figure 9 shows the performance of the proposed model and compares it against the Itti, dynamic, and Fang algorithms. The Itti motion model saliency maps are depicted in column 2, the dynamic model saliency maps in column 3 and the Fang model in column 4. Results obtained using the proposed model are shown in column 5 where from top to bottom, the locally moving snowboarder, flower, and bird are clearly identified as salient objects. Corresponding ground truth frames are shown in column 6, which depict all salient local object movement. Results from our model are subjected

to the presence of significant object motion, which dominates the saliency maps. This is in contrast to the other models where differences between local and global movements are not fully accounted for, therefore, those maps are dominated by spatially attentive features, leading to salient object misclassification. For example, the trees within the background of the snowboard sequence are estimated as an attentive region when a man is performing acrobatics within the frame foreground.

The receiver operating characteristic (ROC) curves and corresponding area under curve (AUC) values, shown in Fig. 10 and the top row in Table 1, respectively, display an objective model evaluation. The results show the proposed method is close to the recent Fang model and exceeds the performance of the Itti motion and dynamic models having 3.5% and 8.2% higher ROC-AUCs, respectively. Further results demonstrating our video VA estimation model across



**Fig. 9** Temporal VAM comparison table: column 1, example original frames from the sequences; column 2, Itti model;<sup>15</sup> column 3, dynamic model;<sup>66</sup> column 4, Fang model;<sup>19</sup> column 5, proposed method, and column 6, ground truth.



**Fig. 10** ROC curve comparing performance of proposed model with state-of-the-art video domain VAMs: Itti model,<sup>15</sup> dynamic model,<sup>66</sup> and Fang model.<sup>19</sup>

**Table 1** AUC and computational times comparing state-of-the-art video domain VAMs.

VA method	Itti <sup>15</sup>	Dynamic <sup>66</sup>	Fang <sup>19</sup>	Proposed
ROC AUC	0.804	0.769	0.867	0.832
Average frame computational time (s)	0.244	0.194	31.54	0.172

four video sequences are shown in Fig. 11. Video saliency becomes more evident when viewed as a sequence rather than from still frames. The video sequences with corresponding saliency maps are available for viewing in Ref. 67.

The bottom row in Table 1 shows the complexity of each algorithm in terms of average frame computational time. The values in the table are calculated from the mean computational time over every frame within the video database and provide the time required to form a saliency map from the original raw frame. All calculations include any transformations required. From the table, the proposed low complex methodology can produce a video saliency map around 30%, 88%, and 0.5% of the time for the Itti, dynamic, and Fang model frames, respectively. Additionally, the proposed model uses the same wavelet decomposition scheme used for watermarking. Therefore, overall visual saliency-based watermarking complexity is low compared to all three methods compared in this paper.

## 4.2 Visual Attention-Based Video Watermarking

The proposed VA-based watermarking is agnostic to the watermark embedding methodology. Thus, it can be used on any existing watermarking algorithm. In our experiments, we use the nonblind embedding proposed by Xia et al.<sup>51</sup> and the blind algorithm proposed by Xie and Arce<sup>52</sup> as our reference algorithms.

A series of experimental results are generated for our video watermarking case study as described in Sec. 3,

analyzing both watermark robustness and imperceptibility. Objective and subjective quality evaluation tools are enforced to provide a comprehensive embedding distortion measure. Robustness against H.264/AVC compression<sup>68</sup> is provided, as common video attacks are comprised of platform reformatting and video compression. Since the VA-based watermarking scheme was presented here as a case study of exploitation of the proposed VAM, our main focus of performance evaluation is on the embedding distortion and the robustness performance with respect to compression attacks. Compression attacks are given focus as watermarking algorithms often employ a higher watermarking strength for encountering the compression and requantization attack. In this work, we demonstrate robustness against H.264/AVC compression, for example. The watermarking evaluation results are reported using the four example video sequences (shown in Fig. 11) from the same data set used for VAM evaluation in the previous section.

$\alpha_{\max}$  and  $\alpha_{\min}$  approximating a PSNR of 35 and 40 dB, respectively, are utilized by applying Eqs. (22) and (23). Four scenarios of varied watermark embedding strengths are considered for the VA-based video watermarking evaluation as follows:

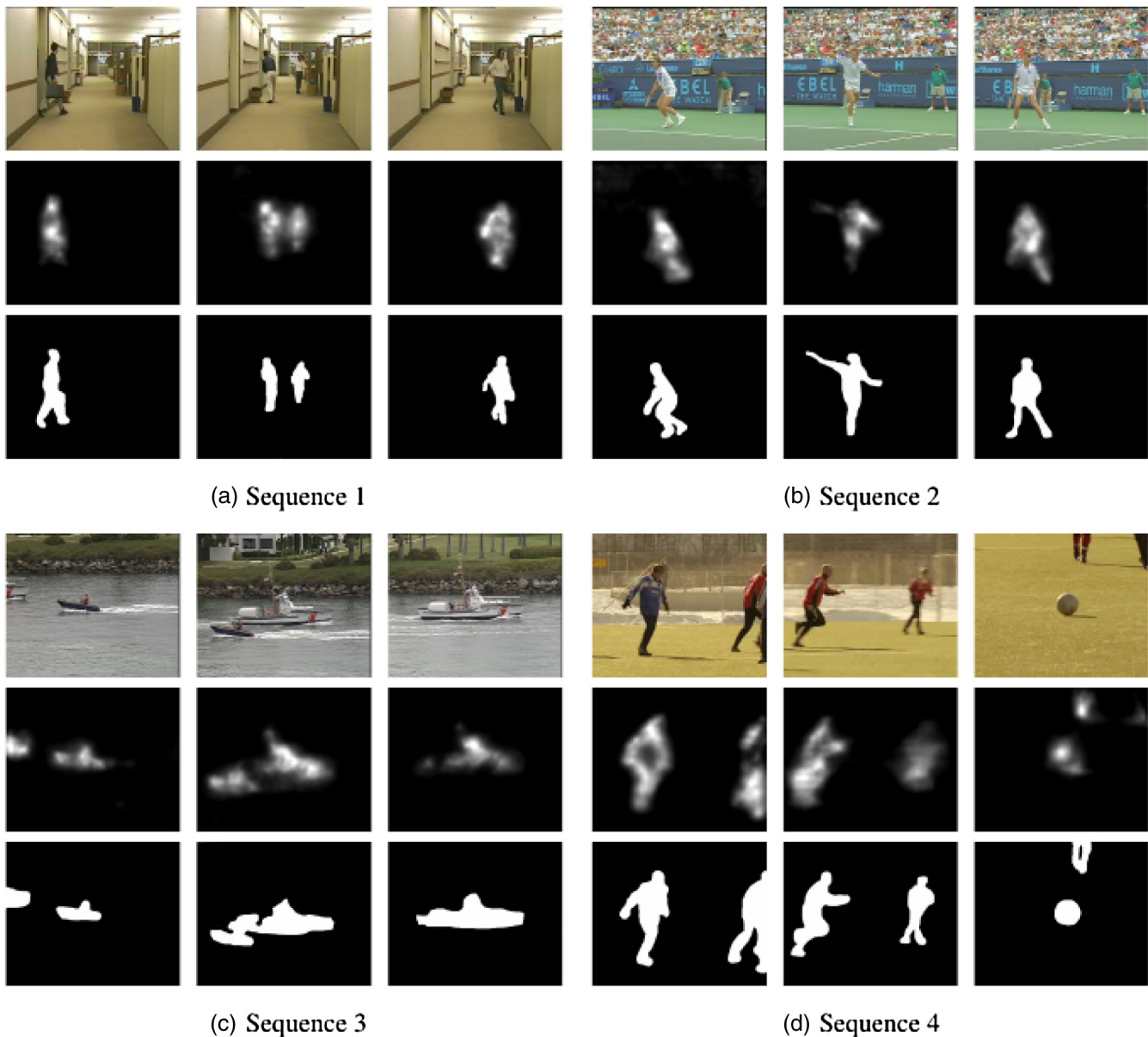
1. a uniform  $\alpha_{\min}$  throughout the entire sequence;
2. the proposed visual VAM-based  $\alpha$  strength;
3. a uniform average watermark strength,  $\alpha_{\text{ave}}$ , chosen as  $\alpha_{\text{ave}} = (\alpha_{\min} + \alpha_{\max})/2$ ; and
4. a uniform  $\alpha_{\max}$  used throughout the entire video sequence.

The experimental results are shown in the following two sections: embedding distortion (visual quality) and robustness.

### 4.2.1 Embedding distortion

The embedding distortion can be evaluated using objective metrics or subjective metrics. While objective quality measurements are mathematical models that are expected to approximate results from subjective assessments and are easy to compute, subjective measurements ensure a viewer's overall opinion of the quality of experience of the visual quality. Often these metrics are complimentary to each other and particularly important in this paper to measure the effect on imperceptibility of the proposed watermark algorithms.

1. "Objective metrics" define a precise value, dependent upon mathematical modeling, to determine visual quality. Such metrics include PSNR, structural similarity index measure (SSIM),<sup>69</sup> just noticeable difference,<sup>70</sup> and video quality metric (VQM).<sup>71</sup> PSNR that calculates the average error between two images is one of the most commonly used visual quality metrics and is described in Eq. (20). Unlike PSNR, SSIM focuses on a quality assessment based on the degradation of structural information. SSIM assumes that the HVS is highly adapted for extracting structural information from a scene. A numeric output is generated between 1 and 0 and higher video quality is represented by values closer to 1. VQM evaluates video quality based upon subjective human perception modeling. It incorporates numerous aspects of early visual processing, including both luma and chroma channels, a combination of temporal and spatial



**Fig. 11** Video visual attention estimation results for four example sequences: row 1, original frame from the sequence; row 2, proposed saliency map; and row 3, ground truth. Video sequences and the VA map sequences are available at Ref. 67.

filtering, light adaptation, spatial frequency, global contrast, and probability summation. A numeric output is generated between 1 and 0 and higher video quality is represented by values closer to 0. VQM is a commonly used video quality assessment metric as it eliminates the need for participants to provide a subjective evaluation.

Although the subjective evaluation is considered as the most suitable evaluation for the proposed method in this paper, the visual quality evaluation in terms of the PSNR, SSIM, and VQM metrics are shown in Tables 2 and 3 for nonblind and blind watermarking schemes, respectively. In both PSNR and SSIM, higher values signify better visual quality. The performance of the four watermarking scenarios in terms of both SSIM and PSNR is rank ordered in terms of the highest visual quality, as follows: low strength embedding ( $\alpha_{\min}$ ) > VA-based algorithm/average strength > high strength embedding ( $\alpha_{\max}$ ). From the tables, PSNR

improvements of  $\sim 3$  dB are achieved when comparing the proposed VA-based approach and constant high strength scenario. The SSIM measures remain consistent for each scenario, with a decrease of 2% for the high-strength watermarking model in most cases. In terms of the VQM metric, which mimics subjective evaluation, the proposed VA-based watermarking consistently performs better than average or high-strength watermarking scenarios.

Objective metrics, such as PSNR, SSIM, and VQM, do not necessarily equal identical perceived visual quality. Two distorted frames with comparable PSNR, SSIM, or VQM metrics do not necessitate coherent media quality. Two independent viewers can undergo entirely different visual experiences, as two similarly distorted frames can provide a contrasting opinion for which contains higher visual quality. To provide a realistic visual quality evaluation, subjective testing is used to analyze the impact of the proposed

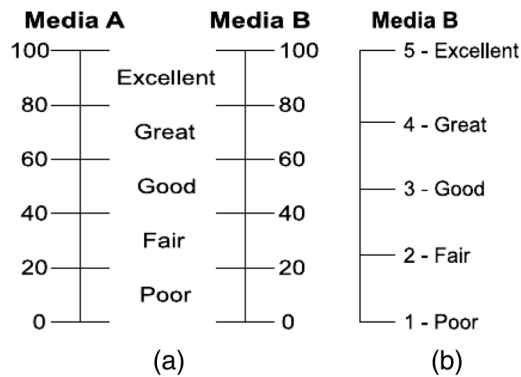


**Table 2** PSNR, SSIM, and VQM average of four video sequences for nonblind watermarking.

	Low strength	Proposed	Average strength	High strength
PSNR	40.15 $\pm$ 0.80	37.39 $\pm$ 0.87	37.47 $\pm$ 0.76	34.93 $\pm$ 0.73
SSIM	0.99 $\pm$ 0.00	0.97 $\pm$ 0.00	0.98 $\pm$ 0.00	0.95 $\pm$ 0.01
VQM	0.10	0.15	0.18	0.25

**Table 3** PSNR, SSIM, and VQM average of four video sequences for blind watermarking.

	Low strength	Proposed	Average strength	High strength
PSNR	40.23 $\pm$ 1.03	36.80 $\pm$ 1.02	37.20 $\pm$ 0.92	34.85 $\pm$ 0.90
SSIM	0.99 $\pm$ 0.00	0.98 $\pm$ 0.00	0.98 $\pm$ 0.00	0.96 $\pm$ 0.01
VQM	0.08	0.13	0.15	0.22

**Fig. 12** Subjective testing visual quality measurement scales (a) DCR continuous measurement scale and (b) ACR ITU 5-point discrete quality scale.

watermarking scheme on the overall perceived human viewing experience.

2. “Subjective evaluation” measures the visual quality by recording the opinion of human subjects on the perceived visual quality. The watermarked videos were viewed by

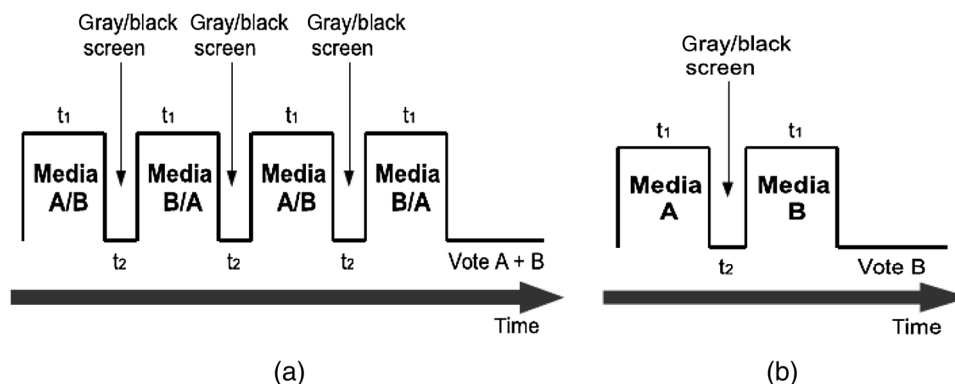
30 subjects, following the standard ITU-T<sup>35</sup> viewing test specifications, often used in compression quality evaluation experiments. The final rating was arrived at by averaging all ratings given by the subjects. This work employs two subjective evaluation metrics that are computed based on the subjective viewing scores, as follows:

“Double stimulus continuous quality test” (DSCQT) subjectively evaluates any media distortion by using a continuous scale. The original and watermarked media is shown to the viewer in a randomized order. The viewer must provide a rating for the media quality of the original and watermarked images individually using a continuous scaling, as shown in Fig. 12(a). Then the degradation category rating (DCR) value is calculated by the absolute difference between the subjective ratings for the two test images.

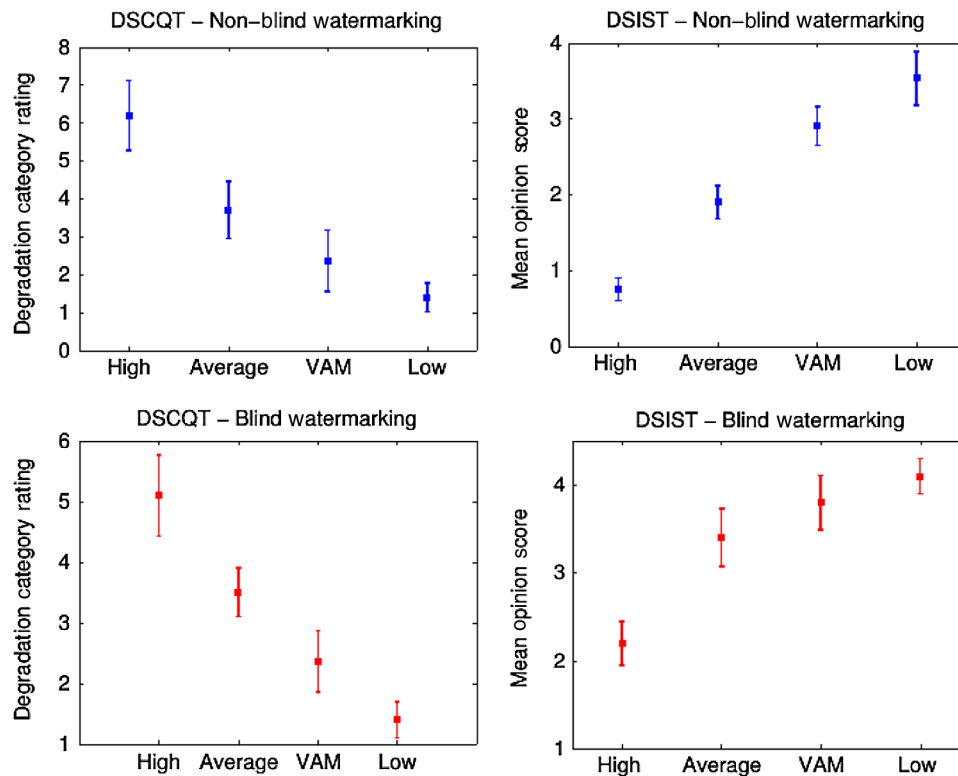
Double stimulus impairment scale test (DSIST) determines the perceived visual degradation between two media sources, A and B, by implementing a discrete scale. A viewer must compare the quality of B with respect to A, on a 5-point discrete absolute category rating (ACR) scale, as shown in Fig. 12(b).

In a subjective evaluation session, first, training images are shown to acclimatize viewers to both ACR and DCR scoring systems. In either of the two subjective tests, a higher value in DCR or ACR scales represents a greater perceived visual quality. Figure 13 shows an overall timing diagram for each subjective testing procedure, showing the sequence of test image displays for scoring by the viewers. Note that the video display time,  $t_1$ , and blank screen time,  $t_2$ , before the change of video, should satisfy the following condition:  $t_1 > t_2$ .

Subjective evaluation performed in this work comprises of DSCQT and DSIST and the results are shown in Fig. 14 for both nonblind and blind watermarking schemes. The top and bottom rows in Fig. 14 show subjective results for the nonblind and blind watermarking cases, respectively, whereas the left and right columns show the results using DSCQT and DSIST evaluation tools. Consistent results are portrayed for both the blind and nonblind scenarios. Figure 14 shows the subjective test results for DCQST and DSIST averaged over four video test sequences. For the DSCQT, the lower the DCR, the better the visual quality, i.e., fewer embedding distortions. In the given results, when comparing the proposed and low strength embedding methodologies, the DCR value only deviates by approximately one unit in the rating scale, suggesting a subjectively similar

**Fig. 13** Stimulus timing diagram for (a) DCR method and (b) ACR method.





**Fig. 14** Subjective video watermarking embedding distortion measures for different embedding scenarios: high strength (high), average strength, VA-based strength selection (VAM) and low strength (low), row 1: nonblind watermarking, row 2: blind watermarking, column 1: DSCQT, and column 2: DSIST.

visual quality. The high-strength watermarking scheme shows a high DCR value indicating significantly higher degradation of subjective visual quality compared with the VAM-based methodology. Similar outcomes are evident from the DSIST plots, where the higher mean opinion score (MOS) on ACR corresponds to better visual quality, i.e., fewer embedding visual distortions. DSIST plots for low-strength and VAM-based schemes show a similar ACR MOS approximately in the range 3 to 4, whereas the high strength watermark yields an ACR of less than 1 for nonblind and nearly 2 for blind watermarking. Compared with an average watermark strength, the proposed watermarking scheme shows an improved subjective image quality in all four graphs by around 0.5 to 1 units. As more data are embedded within the visually salient regions, the subjective visual quality of constant average strength watermarked images is worse than the proposed methodology.

For visual inspection, an example of watermark embedding distortion is shown in Fig. 15. The original, the low strength watermarked, VAM-based watermarked, and the high strength watermarked images are shown in four consecutive columns, where the distortions around the legs of the player with blue jersey (row 1) and around the tennis player (row 2) are distinctively visible in high-strength watermarking.

For each of the blind and nonblind watermarking cases, in both the objective and subjective visual quality evaluations, the low strength watermark and VAM-based watermarking sequences yield similar visual quality, whereas the high strength embedded sequence appears severely more distorted. Low-strength watermarking provides a high imperceptibility but is fragile as discussed in Sec. 4.2.2.

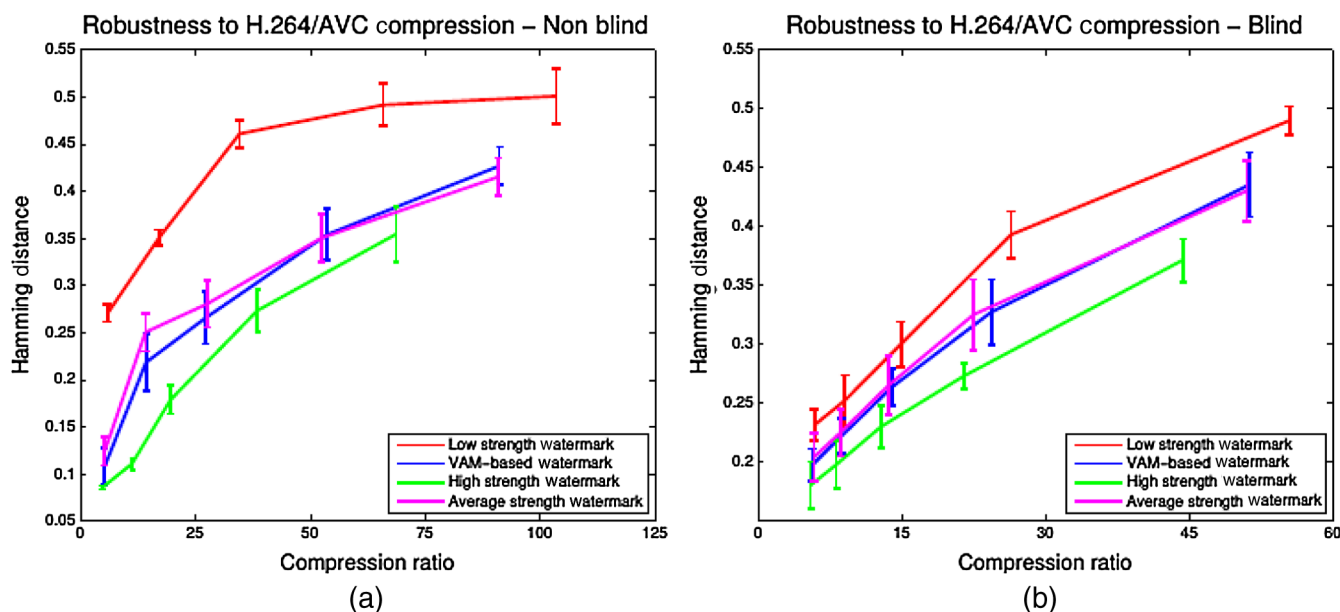
#### 4.2.2 Robustness

Video reformatting and compression are frequent and typically unintentional adversary attacks, hence watermark tolerance for H.264/AVC compression is calculated. Robustness against H.264/AVC compression for both non-blind and blind video watermarking schemes is shown in Figs. 16(a) and 16(b), respectively. For simulating the watermark robustness, five constant quantization parameter (QP) values are implemented to compress the high strength, average strength, VA-based, and low strength test sequences. In both scenarios as shown in the plots, the proposed VA-based methodology shows an increase in robustness compared with the low strength watermark counterpart where a lower Hamming distance indicates better robustness. From the plots in Fig. 16, Hamming distance reductions up to 39% for the nonblind case and 22% for the blind case are possible, when comparing the low and VA-based models. Naturally, the high-strength watermarking scheme portrays a strong Hamming distance but is highly perceptible (low visual quality), as described previously. The proposed watermarking scheme has a slight increased robustness toward H.264/AVC compression, as shown in Fig. 16, when compared against a constant average strength watermark. It is worth noting that for a constant QP value, the compression ratio is inversely proportional to the increase in watermark strength, i.e., as the watermark strength increases, the overall compression ratio decreases due to the extra watermark capacity.

The proposed VA-based method results in a robustness close to the high-strength watermarking scheme, while showing low distortions, as in the low-strength watermarking approach. The incurred increase in robustness coupled with



**Fig. 15** Example frames from soccer and tennis sequences after watermarking with different embedding scenarios (for visual inspection): column 1: original frame, column 2: low strength watermarked frame, column 3: VA-based watermarked frame, and column 4: high strength watermarked frame.



**Fig. 16** Robustness to H.264/AVC compression: average of four video sequences: (a) nonblind watermarking and (b) blind watermarking.

high imperceptibility, verified by subjective and objective metrics, deem the VA-based methodology highly suitable for providing an efficient watermarking scheme.

## 5 Conclusions

In this paper, we have presented a video watermarking algorithm using a motion compensated VAM. The proposed method exploits both spatial and temporal cues for saliency modeled in a motion-compensated spatiotemporal wavelet multiresolution analysis framework. The spatial cues were modeled using the 2-D wavelet coefficients. The temporal cues were modeled using the temporal wavelet coefficients by considering the global and local motion in the video. We have used the proposed VA model in visual-attention-based

video watermarking to achieve robust video watermarking that has minimal or no effect on the visual quality due to watermarking. In the proposed scheme, a two-level watermarking weighting parameter map is generated from the VAM saliency maps using the proposed saliency model and data are embedded into the host image according to the visual attentiveness of each region. By avoiding higher strength watermarking in the visually attentive region, the resulted watermarked video achieved high perceived visual quality while preserving high robustness.

The proposed VAM outperforms the state-of-the-art video VA methods in joint saliency detection and low computational complexity performances. The saliency maps from the proposed method are dominated by the presence of

significant object motion. This is in contrast to the other models where differences between local and global movements are not fully accounted for, therefore, those maps are dominated by spatially attentive features, leading to salient object misclassification. The watermarking performance was verified by performing the subjective evaluation methods as well as the objective metric VQM. For the same embedding distortion, the proposed VA-based watermarking achieved up to 39% (nonblind) and 22% (blind) improvement in robustness against H.264/AVC compression attacks, compared to the existing methodology that does not use the VAM. Finally, the proposed VA-based video watermarking has resulted in visual quality similar to that of low-strength watermarking and robustness similar to those of high-strength watermarking.

### Acknowledgments

We acknowledge the support of the UK Engineering and Physical Research Council (EPSRC), through a Dorothy Hodgkin Postgraduate Award and a Doctoral Training Award at the University of Sheffield.

### References

1. A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.* **12**(1), 97–136 (1980).
2. O. Hikosaka, S. Miyauchi, and S. Shimojo, "Orienting of spatial attention—its reflexive, compensatory, and voluntary mechanisms," *Cognit. Brain Res.* **5**(1–2), 1–9 (1996).
3. L. Itti and C. Koch, "Computational modelling of visual attention," *Nat. Rev. Neurosci.* **2**, 194–203 (2001).
4. R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annu. Rev. Neurosci.* **18**(1), 193–222 (1995).
5. D. Ress, B. T. Backus, and D. J. Heeger, "Activity in primary visual cortex predicts performance in a visual detection task," *Nat. Neurosci.* **3**(9), 940–945 (2000).
6. J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nat. Rev. Neurosci.* **5**(6), 495–501 (2004).
7. R. Achanta et al., "Frequency-tuned salient region detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1597–1604 (2009).
8. A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 185–207 (2013).
9. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 1254–1259 (1998).
10. L. ZhiQiang, F. Tao, and H. Hong, "A saliency model based on wavelet transform and visual attention," *Inf. Sci.* **53**, 738–751 (2010).
11. S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2376–2383 (2010).
12. B. P. Olveczky, S. A. Baccus, and M. Meister, "Segregation of object and background motion in the retina," *Nature* **423**(6938), 401–408 (2003).
13. F. Guraya and F. Cheikh, "Predictive visual saliency model for surveillance video," in *Proc. IEEE European Signal Processing Conf. (EUSIPCO '11)*, pp. 554–558 (2011).
14. O. Meur, P. Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Res.* **47**, 2483–2498 (2007).
15. L. Itti and N. Dhavale, "Realistic avatar eye and head animation using a neurobiological model of visual attention," *Proc. SPIE* **5200**, 64 (2004).
16. Y. Tong et al., "A spatiotemporal saliency model for video surveillance," *Cognit. Comput.* **3**, 241–263 (2011).
17. C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.* **19**, 185–198 (2010).
18. S.-H. Zhong et al., "Video saliency detection via dynamic consistent spatio-temporal attention modelling," in *Association for the Advancement of Artificial Intelligence* (2013).
19. Y. Fang et al., "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Trans. Image Process.* **23**, 3910–3921 (2014).
20. Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM Int. Conf. on Multimedia*, pp. 815–824 (2006).
21. M. Dorr, E. Vig, and E. Barth, "Colour saliency on video," in *Bio-Inspired Models of Network, Information, and Computing Systems*, Vol. 87, pp. 601–606 (2012).
22. C. C. Loy, X. Tao, and G. S. Gong, "Salient motion detection in crowded scenes," in *Int. Symp. on Communications Control and Signal Processing, (ISCCSP '12)*, pp. 1–4 (2012).
23. D. Que et al., "An ROI image watermarking algorithm based on lifting wavelet transform," in *Proc. Int. Conf. on Signal Processing*, Vol. 4, pp. 16–20 (2006).
24. R. Ni and Q. Ruan, "Region of interest watermarking based on fractal dimension," in *Proc. Int. Conf. on Pattern Recognition*, pp. 934–937 (2006).
25. R. Wang, Q. Cheng, and T. Huang, "Identify regions of interest (ROI) for video watermark embedment with principle component analysis," in *Proc. ACM Int. Conf. on Multimedia*, pp. 459–461 (2000).
26. C. Yiping et al., "Region of interest fragile watermarking for image authentication," in *Int. Multi-Symp. on Computer and Computational Sciences (IMSCCS '06)*, Vol. 1, pp. 726–731 (2006).
27. L. Tian et al., "An integrated visual saliency-based watermarking approach for synchronous image authentication and copyright protection," *Image Commun.* **26**, 427–437 (2011).
28. B. Ma et al., "Salient region detection for biometric watermarking," in *Computer Vision for Multimedia Applications: Methods and Solutions*, pp. 218–236, IGI Global (2010).
29. A. Sur et al., "A new image watermarking scheme using saliency based visual attention model," in *Annual IEEE India Conf. (INDICON '09)*, pp. 1–4 (2009).
30. M. Oakes, D. Bhowmik, and C. Abhayaratne, "Visual attention-based watermarking," in *IEEE Int. Symp. on Circuits and Systems (ISCAS '11)*, pp. 2653–2656 (2011).
31. D. Bhowmik, M. Oakes, and C. Abhayaratne, "Visual attention-based image watermarking," *IEEE Access* **4**, 8002–8018 (2016).
32. M. Oakes, "Attention driven solutions for robust digital watermarking within media," PhD Thesis, University of Sheffield (2014).
33. L. Tian et al., "Authentication and copyright protection watermarking scheme for h. 264 based on visual saliency and secret sharing," *Multimedia Tools Appl.* **74**(9), 2991–3011 (2015).
34. A. Barari and S. V. Dhavale, "Video saliency detection for visual cryptography-based watermarking," in *Innovative Research in Attention Modeling and Computer Vision Applications*, p. 132, IGI Global (2015).
35. H. G. Koumaras, "Subjective video quality assessment methods for multimedia applications," Technical Report ITU-R BT.500-11, Geneva, Switzerland (2008).
36. M. Oakes and C. Abhayaratne, "Visual saliency estimation for video," in *Proc. 13th Int. Workshop on Image Analysis for Multimedia Interactive Services*, pp. 1–4 (2012).
37. H. Wilson, "Psychophysical models of spatial vision and hyper-acuity," *Spat. Vision* **10**, 64–81 (1991).
38. R. Jin, Y. Qi, and A. Hauptmann, "A probabilistic model for camera zoom detection," in *Proc. Int. Conf. on Pattern Recognition*, Vol. 3, pp. 859–862 (2002).
39. H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: speeded up robust features," in *European Conf. on Computer Vision*, pp. 404–417 (2006).
40. S. Choi and J. W. Woods, "Motion-compensated 3-D subband coding of video," *IEEE Trans. Image Process.* **8**(2), 155–167 (1999).
41. Y. Andreopoulos et al., "In-band motion compensated temporal filtering," *Signal Process. Image Commun.* **19**(7), 653–673 (2004).
42. D. Bhowmik and C. Abhayaratne, "2D+T wavelet domain video watermarking," *Adv. Multimedia* **2012**, 973418 (2012).
43. D. Mahapatra, S. Winkler, and S. Yen, "Motion saliency outweighs other low-level features while watching videos," *Proc. SPIE* **6806**, 68060P (2008).
44. D. S. Taubman and M. W. Marcellin, *JPEG2000 Image Compression Fundamentals, Standards and Practice*, Springer (2002).
45. P. Chen and J. W. Woods, "Bidirectional MC-EZBC with lifting implementation," *IEEE Trans. Circuits Syst. Video Technol.* **14**(10), 1183–1194 (2004).
46. G. Bhatnagar, Q. M. J. Wu, and B. Raman, "Robust gray-scale logo watermarking in wavelet domain," *Comput. Electr. Eng.* **38**, 5, 1164–1176 (2012).
47. A. Piper, R. Safavi-Naini, and A. Mertins, "Resolution and quality scalable spread spectrum image watermarking," in *Proc. 7th Workshop on Multimedia and Security: MM&Sec'05*, pp. 79–90 (2005).
48. D. Bhowmik and C. Abhayaratne, "A generalised model for distortion performance analysis of wavelet based watermarking," *Lect. Notes Comp. Sci.* **5450**, 363–378 (2008).
49. C. Abhayaratne and D. Bhowmik, "Scalable watermark extraction for real-time authentication of JPEG2000 images," *J. Real-Time Image Process.* **6**(4), 19 (2011).
50. D. Bhowmik and C. Abhayaratne, "Quality scalability aware watermarking for visual content," *IEEE Trans. Image Process.* **25**(11), 5158–5172 (2016).
51. X. Xia, C. G. Boncellet, and G. R. Arce, "Wavelet transform based watermark for digital images," *Opt. Express* **3**, 497–511 (1998).



52. L. Xie and G. R. Arce, "Joint wavelet compression and authentication watermarking," in *Proc. IEEE Int. Conf. on Image Processing*, Vol. 2, pp. 427–431 (1998).
53. M. Barni, F. Bartolini, and A. Piva, "Improved wavelet-based watermarking through pixel-wise masking," *IEEE Trans. Image Process.* **10**, 783–791 (2001).
54. C. Jin and J. Peng, "A robust wavelet-based blind digital watermarking algorithm," *Int. J. Inf. Technol.* **5**(2), 358–363 (2006).
55. R. S. Shekhawat, V. S. Rao, and V. K. Srivastava, "A biorthogonal wavelet transform based robust watermarking scheme," in *Proc. IEEE Conf. on Electrical, Electronics and Computer Science (SCECS)*, pp. 1–4 (2012).
56. D. Bhowmik and C. Abhayaratne, "On robustness against JPEG2000: a performance evaluation of wavelet-based watermarking techniques," *Multimedia Syst.* **20**(2), 239–252 (2014).
57. Q. Gong and H. Shen, "Toward blind logo watermarking in JPEG-compressed images," in *Sixth Int. Conf. on Parallel and Distributed Computing, Applications and Technologies (PDCAT '05)*, pp. 1058–1062 (2005).
58. J. R. Kim and Y. S. Moon, "A robust wavelet-based digital watermarking using level-adaptive thresholding," in *Proc. Int. Conf. on Image Processing (ICIP '99)*, Vol. 2, pp. 226–230 (1999).
59. V. Saxena, M. Gupta, and D. T. Gupta, "A wavelet-based watermarking scheme for color images," *IUP J. Telecommun.* **5**, 56–66 (2013).
60. F. Huo and X. Gao, "A wavelet based image watermarking scheme," in *Proc. Int. Conf. on Image Processing (ICIP '06)*, 2573–2576 (2006).
61. D. Kundur and D. Hatzinakos, "Digital watermarking using multiresolution wavelet decomposition," in *Proc. of the 1998 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 5, pp. 2969–2972 (1998).
62. V. S. Verma and J. R. Kumar, "Improved watermarking technique based on significant difference of lifting wavelet coefficients," *Signal Image Video Process.* **9**(6), 1443–1450 (2015).
63. P. Meerwald, "Quantization watermarking in the JPEG2000 coding pipeline," in *Proc. Int. Working Conf. on Communications and Multimedia Security*, pp. 69–79 (2001).
64. D. Aggarwal and K. S. Dhindsa, "Effect of embedding watermark on compression of the digital images," *Comput. Res. Repository* **1002**, 5–8 (2010). <https://arxiv.org/abs/1002.3984>
65. K. Fukuchi et al., "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, pp. 638–641 (2009).
66. X. Hou and L. Zhang, "Dynamic visual attention: searching for coding length increments," in *Advances in Neural Information Processing Systems*, pp. 681–688 (2008).
67. M. Oakes and C. Abhayaratne, "Global Motion Compensated Video Saliency Estimation," (accessed 15 Oct 2016).
68. H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.* **17**(9), 1103–1120 (2007).
69. Z. Wang et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.* **13**, 600–612 (2004).
70. A. B. Watson, "Visual optimization of dct quantization matrices for individual images," in *American Institute of Aeronautics and Astronautics (AIAA)*, Vol. 9, pp. 286–291 (1993).
71. A. B. Watson, J. Hu, and J. F. McGowan III, "DVQ: a digital video quality metric based on human vision," *J. Electron. Imaging* **10**, 20–29 (2001).

**Matthew Oakes** graduated with an MEng degree in electronic and electrical engineering from the University of Sheffield in 2009. He received his PhD in electronic and electrical engineering also at the University of Sheffield in 2014. He is currently working at the University of Buckingham as a knowledge transfer partnership associate. His main expertise lies in image/video processing, compression, digital watermarking, and visual saliency estimation. His current research includes biometric security systems and machine learning.

**Deepayan Bhowmik** received his PhD in electronic and electrical engineering from the University of Sheffield, UK, in 2011. Previously, he worked as a research associate at Heriot-Watt University, Edinburgh, UK and the University of Sheffield, UK, and a system engineer in ABB Ltd., India. He is currently working as a lecturer in Sheffield Hallam University, Sheffield, UK. His current research interests include computer vision, machine learning, embedded imaging hardware on FPGA, and multimedia security.

**Charith Abhayaratne** received his BE degree in electrical and electronic engineering from the University of Adelaide, Australia, in 1998, and his PhD in the same from the University of Bath, UK, in 2002. He is currently a lecturer in the Department of Electronic and Electrical Engineering at the University of Sheffield, UK. His research interests include video and image compression, watermarking, image and video analysis, multidimensional signal processing, graph spectral analysis, and computer vision.