Sheffield Hallam University

Global motion compensated visual attention-based video watermarking

OAKES, Matthew, BHOWMIK, Deepayan <http://orcid.org/0000-0003-1762-1578> and ABHAYARATNE, Charith

Available from Sheffield Hallam University Research Archive (SHURA) at:

https://shura.shu.ac.uk/14390/

This document is the Accepted Version [AM]

Citation:

OAKES, Matthew, BHOWMIK, Deepayan and ABHAYARATNE, Charith (2016). Global motion compensated visual attention-based video watermarking. Journal of Electronic Imaging, 25 (6), 061624. [Article]

Copyright and re-use policy

See http://shura.shu.ac.uk/information.html

Sheffield Hallam University Research Archive
http://shura.shu.ac.uk

Global Motion Compensated Visual Attention based Video Watermarking

Matthew Oakes[a], **Deepayan Bhowmik**[b,*], **Charith Abhayaratne**[c]
[a]University of Buckingham, Buckingham, United Kingdom, MK18 1EG.
[b]Department of Computing, Sheffield Hallam University, Sheffield, United Kingdom, S1 1WB.
[c]Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, United Kingdom, S1 3JD.

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 suffers from embedding distortions resulting in poor visual quality in host media. This paper proposes a novel 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 modelling uses the spatial wavelet coefficients while the temporal modelling accounts for both local and global motion to arrive at the spatio-temporal visual attention model for video. The model is then used to develop a new video watermarking algorithm, where a two-level watermarking weighting parameter map is generated from the VAM saliency maps using the saliency model and data is embedded into the host image according to the visual attentiveness of each region. By avoiding higher strength watermarking in visually attentive region, the resulted 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 performances. For the same embedding distortion, the proposed visual attention based watermarking achieves up to 39% (non-blind) and 22% (blind) improvement in robustness against H.264/AVC compression, compared to the 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 robustness similar to those of high-strength watermarking.

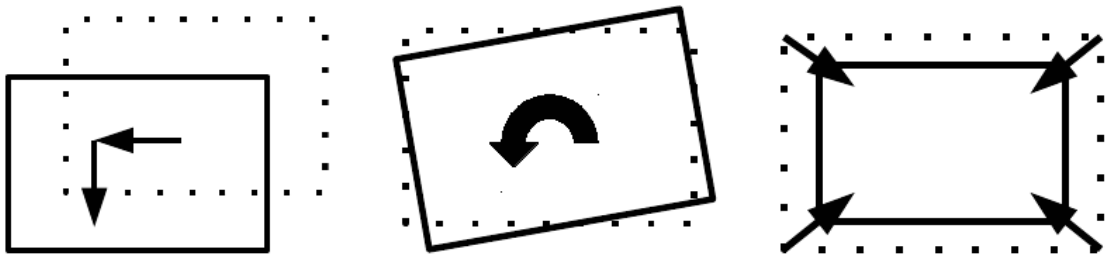
Keywords: Video visual attention model, motion compensation, video watermarking, robustness, subjective visual quality evaluation.

*Deepayan Bhowmik, deepayan.bhowmik@shu.ac.uk

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 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 and hence 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.

The human visual system (HVS) is sensitive to many features which lead to attention being drawn towards specific regions in a scene and a well studied topic in psychology and biology.[1,2] Visual attention (VA) is an important and complex biological process that helps to identify potential danger, *e.g.*, prey, predators quickly in a cluttered visual world[3] as attention to one target leaves other targets less available.[4] Recently a considerable effort was noticed in the literature in modelling visual attention[5] that has applications in many related domains including media quality evaluation. Areas of visual interest stimulate neural nerve cells, causing human gaze to fixate



(a) Camera Panning (b) Camera Tilting (c) Camera Zooming

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

towards a particular scene area. The Visual Attention Model (VAM) highlights these visually sensitive regions, which stimulate a neural response within the primary visual cortex.[6] Whether that neural vitalization be from contrast in intensity, a distinctive face, unorthodox motion or a dominant colour, these stimulative regions diverge human attention providing highly useful saliency maps within the media processing domain.

Human vision behavioural studies[7] and feature integration theory[8] have prioritised the combination of three visually stimulating low level features: intensity, colour and orientation which comprise the concrete foundations for numerous image domain saliency models.[3,9,10] Most saliency models often use Multi-Resolution Analysis (MRA).[11–13] Temporal features must be considered as moving objects are more eye-catching than most static locations.[14] Seldom work has been directed towards 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[3,15–20] exist as a supplementary extension from their image domain algorithms. Research estimating VA within video can also be derived from exploiting spatiotemporal cues,[21,22] structural tensors[23] and optical flow.[24]

However none of these algorithms explicitly capture the spatio-temporal 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 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. Fig 1 shows three causes for global motion. This paper proposes a new video visual attention model that accounts 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 novel 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 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 Section 2).

Related work includes defining a Region of Interest (ROI)[25–30] and increasing the watermark strength in the ROI to address cropping attacks. However, in these works, the ROI extraction were 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 towards any additional embedding artefacts, and *b)* scenes exhibiting sparse salience will potentially contain extensively fragile or no watermark data. Moreover, Sur *et al.*[31] proposed a pixel domain algorithm to improve embedding distortion using an existing visual saliency model described in.[3] However, the algorithm only discusses its limited observation on perceptual quality without considering any robustness.

Our previous work[32,33] 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 analyse each individual frame. Human attention will naturally be diverged towards 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 visual attention based image and video watermarking was first introduced in our early work.[32,34] Recent work following this concept can be found in watermarking H.264 video[35] and application on cryptography.[36] On the contrary, in this paper, we propose a video watermark embedding strategy based on visual attention modelling that uses the same spatio-temporal decomposition used in the video watermarking scheme. In addition, the visual attention model compensate 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 comparing with existing schemes. Subjective tests for media quality assessment recommended by the International Telecommunication Union (ITU),[37] largely missing in the watermarking literature, are also conducted to complement the objective measurements. Major contributions of this paper are:

- A new motion compensated spatio-temporal video visual attention model that considers object motion between frames as well as global motions due to camera movement.
- New blind and non-blind video watermarking algorithms that is highly imperceptible and robust against compression attacks.
- Subjective tests that evaluate visual quality of the proposed watermarking algorithms.

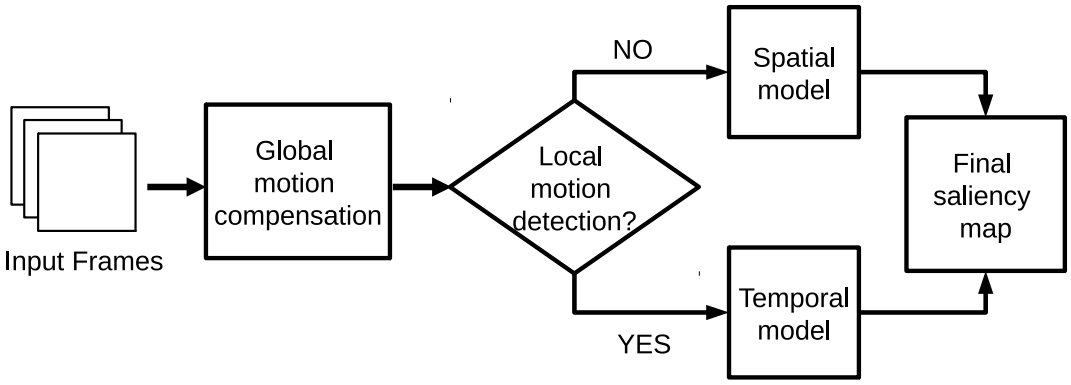


Fig 2: Proposed video visual attention model functional block diagram.

The saliency model and the watermarking algorithms are evaluated using existing video datasets described in Section 4.1. Initial concept of the motion compensated video attention model was reported earlier in the form of a conference publication[38] while this paper discusses the proposed scheme in details 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 novel method is proposed to detect any saliency information within a video. The proposed methods incorporate motion compensated spatio-temporal wavelet decomposition combined with HVS modelling to capture any saliency information. A unique approach combining salient temporal, intensity, colour, orientation contrasts formulate the essential video saliency methodology.

Physiological and psychophysical evidence demonstrate visually stimulating regions occur at different scales within media[39] and the object motion within the scene.[14] Consequently, models proposed in this work exploit the identifiable multi-resolution property of the wavelet transform that incorporates a motion compensation algorithm to generate the model. By exploiting the multi resolution spatio-temporal representation of the wavelet transform, VA is estimated directly from within the wavelet domain. The video saliency model is divided into three subsections. Firstly, Section 2.1 describes the global motion compensation following the description of the spatial saliency model in Section 2.2 and Section 2.3 illustrates the temporal saliency feature map generation. Finally, Section 2.4 combines spatio-temporal model to estimate video visual saliency. An over all functional block diagram of our proposed model is shown in Fig 2. For the spatial saliency model in this work, we adopted our image visual attention model proposed in 32, 33.

2.1 Global Motion Compensated Frame Difference

Compensation for global motion is dependant upon homogeneous Motion Vector (MV) detection, consistent throughout the frame. Fig 3 considers the motion estimation between two consecutive frames, taken from the coastguard sequence. A fixed block size based upon the frame resolution determines 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 a MV of

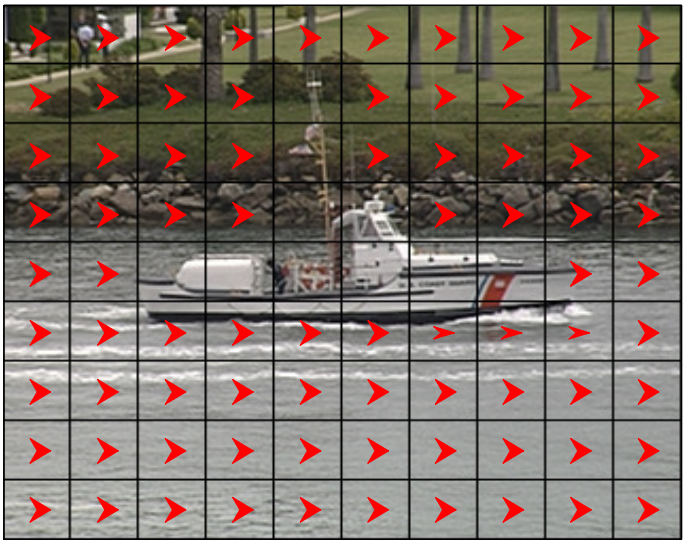


Fig 3: Motion block estimation

zero. Firstly, it is assumed there is a greater percentage of pixels within moving objects than the background, so large densities of comparative MVs are the result from dynamic camera action. To compensate for camera panning, the entire reference frame is spatially translated by the most frequent MV, the global camera MV, $\overrightarrow{M}_{global}$. This process is applied, prior to the 2D+t wavelet decomposition to deduce global motion compensated saliency estimation. The global motion compensation is described in Eq.(1):

$$\overrightarrow{M}_{object} = \overrightarrow{M}_{total} - \overrightarrow{M}_{global}, \tag{1}$$

where $\overrightarrow{M}_{object}$ is the local object MV and $\overrightarrow{M}_{total}$ 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.[40] Speeded Up Robust Features (SURF) detection[41] 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 colour image / frame, RGB colour space is converted to YUV colour spectral space as the latter exhibits prominent intensity variations through its luminance channel Y. Firstly, the 2D forward DWT (FDWT) is applied on each Y, U and V channel to decompose them in multiple levels. The 2D 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. DWT captures horizontal, vertical

and diagonal contrasts within an image, respectively, portraying prominent edges in various orientations. Due to the dyadic nature of the multi-resolution 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 normalised so that the overall saliency contributions from each subband and prevents biassing towards 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 illustrated in Eq. (2):

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

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

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

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

where $\overline{lh}_X$, $\overline{hl}_X$ and $\overline{hh}_X$ are the normalised set of subband feature maps.

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

$$S = \sum_{\forall X \in \{Y,U,V\}} w_X * S_X, \tag{4}$$

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:

$$S_X = \overline{LH}_X + \overline{HL}_X + \overline{HH}_X. \tag{5}$$

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

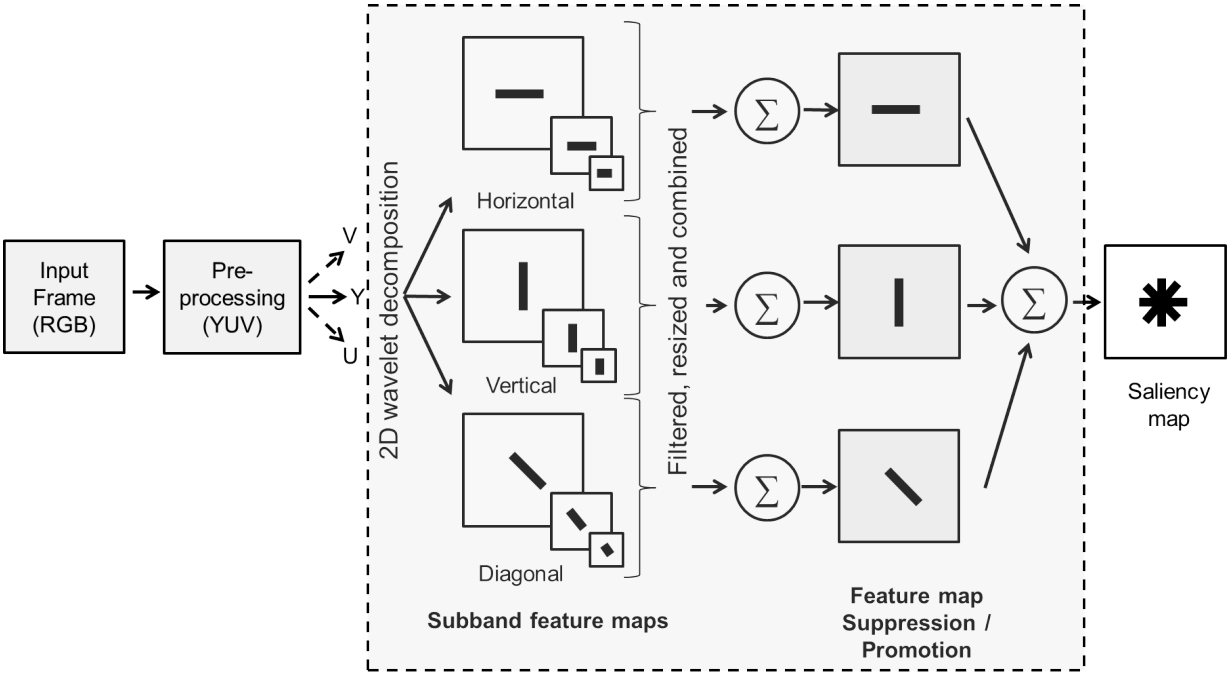


Fig 4: Overall functional diagram of the spatial visual saliency model.

2.3 Temporal Saliency Model

2.3.1 2D+t Wavelet Domain

We extend our spatial saliency model towards video domain saliency, logically by uitlising a 3D wavelet transform. Video coding research provides evidence that differing texture and motion characteristics occur after wavelet decomposition from the t+2D domain[42] and incorporating its alternative technique, the 2D+t transform.[43,44] The t+2D 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, 2D+t decomposition produces greater transform energy within the higher frequency components, *i.e.*, a greater amounts of larger and non-zero coefficients and reduces computational complexity to a great extent. A description on reduced computational complexity by using 2D+t compared to t+2D can be found in 44. Therefore, in this work we have used a 2D+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 Section 2.2 constitutes the spatial element for the video saliency model where as this section concentrates upon establishing temporal saliency maps, S_{Temp}.

Similar methodology to expose temporal conspicuousness is implemented in comparison to the spatial model in Section 2.2. Firstly, the existence of any palpable local object motion is de-

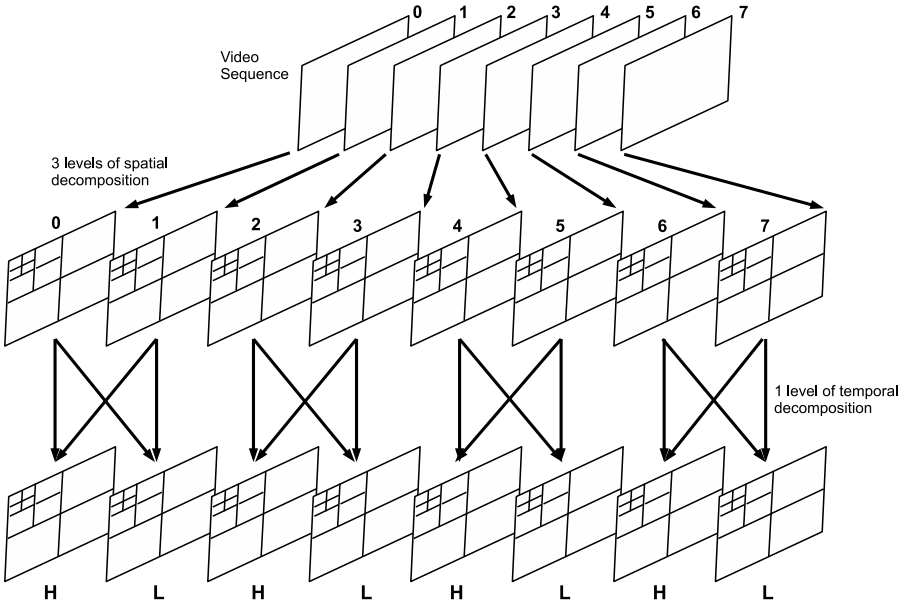


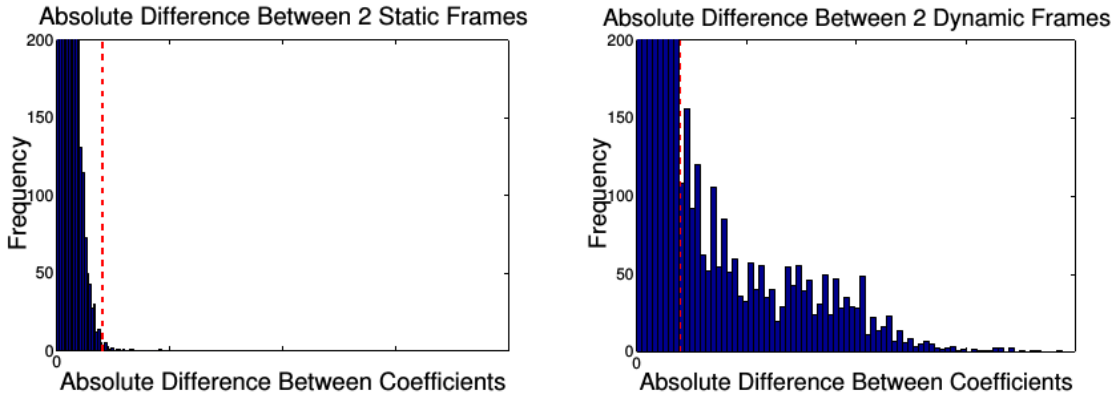
Fig 5: 2D+t wavelet decomposition.

termined within the sequence. Fig 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 Section 2.1). Change within lighting, noise and global motion compensation error account for the peaks present within Fig 6a), whereas the contribution from object movement is also present within Fig 6b). 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:

$$D(x, y) = |F_1(x, y) - F_2(x, y)|. \tag{6}$$

From the histograms shown within Fig 6(a) and Fig 6(b), a local threshold value of $T = D_{max}/10$ determines motion classification, where D_{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_{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 2D+t wavelet decomposition, which is shown in Fig 5. The decomposed data is forged using comparable logic as Eq.(2), as all transformed coefficients are segregated into 1 of 3 temporal



(a) (b)

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

subband feature maps. This process is described in Eq.(7):

$$\begin{aligned} LHt_t &= \sum_{i=1}^{n}(|LHt_i|^{\uparrow 2^i} * \tau_i), \\ HLt_t &= \sum_{i=1}^{n}(|HLt_i|^{\uparrow 2^i} * \tau_i), \\ HHt_t &= \sum_{i=1}^{n}(|HHt_i|^{\uparrow 2^i} * \tau_i), \end{aligned} \tag{7}$$

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 both 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 Eq.(8):

$$S_{Temp} = LHt_t + HLt_t + HHt_t, \tag{8}$$

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, γ, determined from Eq. (6) is calculated in Eq.(9):

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

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.[45] Consequently, if no local motion is detected with a frame, the spatial model contributes towards the final saliency map in its entirety,

hence γ is a binary variable. The equation forging the overall saliency map is,

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

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 a novel algorithm that provides a solution towards blind and non-blind VA-based video watermarking. The video saliency model described in Section 2 is utilized within the video watermarking framework to determine the watermarking embedding strength. Coinciding with the previous video VA model, watermark data is embedded within the 2D+t wavelet domain as outlined in Section 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 artefacts 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 non-blind) 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 α in salient regions and non-salient regions, respectively, ensures higher imperceptibility of the watermarked image distortions while keeping greater robustness.

3.1 The 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 favoured in the current research era. The wavelet domain is also compliant within many image coding, *e.g.*, JPEG2000[46] and video coding, *e.g.*, Motion JPEG2000, Motion-Compensated Embedded Zeroblock Coding (MC-EZBC),[47] schemes, leading to smooth adaptability within modern frameworks. Due to the multi-resolution decomposition and the property to retain spatial synchronisation, which are not provided by other transforms (the Discrete Cosine Transform (DCT) for example), the Discrete Wavelet Transform (DWT) provides an ideal choice for robust watermarking.[48–57]

The FDWT is applied on the host image before watermark data is embedded within the selected subband coefficients. The Inverse Discrete Wavelet Transform (IDWT) reconstructs the watermarked image. The extraction operation is performed after the FDWT. The extracted watermark data is 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 58.

3.1.1 Non-blind Watermarking

Magnitude-based multiplicative watermarking[32,53,55,59–61] is a popular choice when using a non-blind watermarking system, due to its simplicity. Wavelet coefficients are modified based on the watermark strength parameter, α, 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:

$$C'(m,n) = C(m,n) + \alpha W(m,n) C(m,n). \tag{11}$$

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

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

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

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

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

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

3.1.2 Blind Watermarking

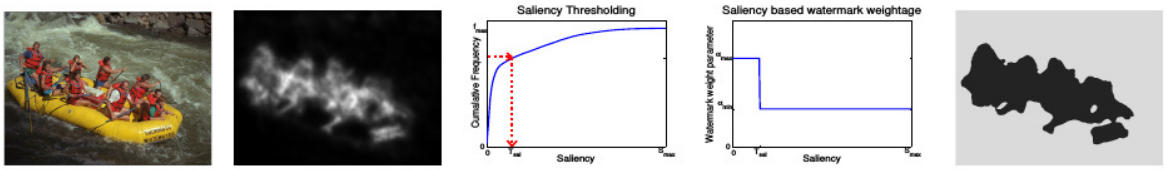
Quantization-based watermarking[54,62–66] is a blind scheme which relies on modifying various coefficients towards a specific quantization step. As proposed in 54, the algorithm is based on modifying the median coefficient towards the step size, δ, by using a running non-overlapping 3×1 window. The altered coefficient must retain the median value of the three coefficients within the window, after the modification. The equation calculating δ is described as follows:

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

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

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

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



(a) (b) (c) (d) (e)

Fig 7: (a) Example host image (b) VAM saliency map (saliency is proportional to the grey scale) (c) Cumulative saliency histogram (d) α step graph (e) α strength map (dark corresponds to low strength).

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:

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

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 subsection presents the threshold-based saliency map segmentation which is used for adapting the watermarking algorithms described in Section 3.1 in order to change the watermark strength according to the underlying visual attention properties. Fig 7(a) and Fig 7(b) show an example original host frame and its corresponding saliency map, respectively, generated from the proposed methodology in Section 2. In Fig 7(b), the light and dark regions, within the saliency map, represent the visually attentive and non-attentive areas, respectively. At this point, we employ thresholding to quantise 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 non-blind watermarking schemes, in Section 3.1, the host media source is only available within non-blind 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 quantised 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 Section 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, firstly, 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

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

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, α, in Eq. (11) and Eq. (15) is made variable $\alpha(j,k)$, dependant upon T_s, as follows:

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

where $\alpha(j,k)$ is the adaptive watermark strength map giving the α value for a the corresponding saliency at a given pixel coordinate (j,k). The watermark weighting parameters, α_{min} and α_{max} correspond to the high and low strength, values respectively and their typical values are determined from the analysis within Section 3.3. As shown in Fig 7(d), the most and the least salient regions are given watermark weighting parameters of α_{min} and α_{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 α.

3.3 Watermark Embedding Strength Calculation

The watermark weighting parameter strengths, α_{max} and α_{min} can be calculated from the visible artifact PSNR limitations within the image. Visual distortion becomes gradually noticeable as the overall Peak Signal to Noise Ratio (PSNR) drops below $40 \sim 35$dB,[67] so minimum and maximum PSNR requirements are set to approximate 35dB and 40dB, respectively, for both the blind and non-blind watermarking schemes. These PSNR limits ensure 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 α.

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

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

where M is the maximum coefficient value of the data, $I(j,k)$ and $I'(j,k)$ is 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.[50] Therefore, Eq.(20) can be redefined on transform domain for non-blind magnitude based multiplicative watermarking, shown

in Eq.(11), as follows:

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

By rearranging for α, an expression determining the watermark weighting parameter, depending on the desired PSNR value is derived for non-blind watermarking in Eq.(22) as follows:

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

Similarly for the blind watermarking scheme described in Section 3.1.2, PSNR in transform domain can be estimated by substituting the median and modified median coefficients, $C_{(med)}$ and $C'_{(med)}$, 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:

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

Eq.(23) determines the total coefficient modification for a given PSNR requirement, hence is used to α in Eq.(15).

3.4 Saliency Map Reconstruction

For non-blind watermarking, the host data is 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 got pixel values slightly different to the original host media. Thresholding the saliency map into 2 levels, as described in Section 3.2, ensures high accuracy within the saliency model reconstruction for blind watermarking. Further experimental objective analysis reveals that use of thresholding improves the saliency coefficients match upto 99.4% compared to approximately only 55.6% of coefficients when thresholding was not used and hence reconstruction errors are greatly reduced.

4 Experimental Results and Discussion

The performance of the proposed video visual attention 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 Section 4.1. The video visual attention based watermarking is evaluated in terms of embedding distortion and robustness to compression in Section 4.2.

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

4.1 Visual Attention Model Evaluation

For attention model evaluation, the video dataset is taken from the literature 68, which comprises 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 2D 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*,[17] *Dynamic*[69] and *Fang*[21] video visual attention models, 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 dependant upon locating energy peaks within incremental length coding. A more recent *Fang* algorithm uses a spatiotemporally adaptive entropy-based uncertainty weighting approach.

Fig 9 shows the performance of the proposed model and compares 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 movement are not fully accounted for and 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.

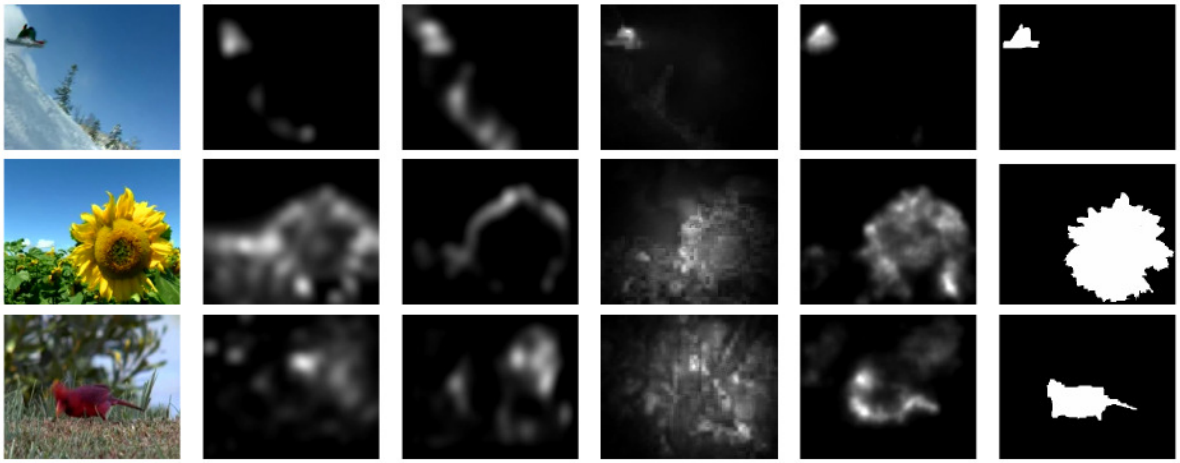


Fig 9: Temporal visual attention model comparison table: *column 1* - example original frames from the sequences; *column 2* - *Itti* model;[17] *column 3* - *Dynamic* model;[69] *column 4* - *Fang* model;[21] *Column 5* - proposed method and *column 6* - ground truth.

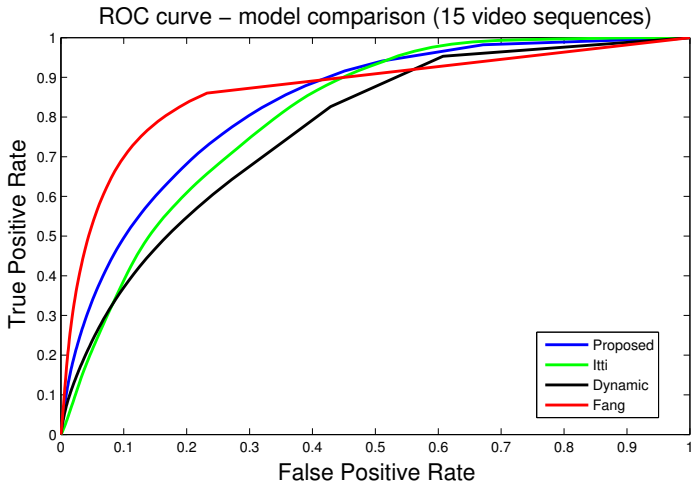


Fig 10: ROC curve comparing performance of proposed model with state-of-the-art video domain visual attention models: *Itti* model,[17] *Dynamic* model[69] and *Fang* model.[21]

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 recent hFang model and exceeds the performance of the *Itti* motion and *Dynamic* models having a 3.5% and 8.2% higher ROC-AUC, respectively. Further results demonstrating our video visual attention estimation model across 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 at the following website address: http://svc.group.shef.ac.uk/VAvideo.html.

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%

Table 1: AUC and Computational time comparing state-of-the-art video domain visual attention models.

Visual Attention Method	Itti[17]	Dynamic[69]	Fang[21]	Proposed
ROC AUC	0.804	0.769	0.867	0.832
Average frame computational time (sec)	0.244	0.194	31.54	0.172

of the time for an *Itti*, *Dynamic* and *Fang* model frame, 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

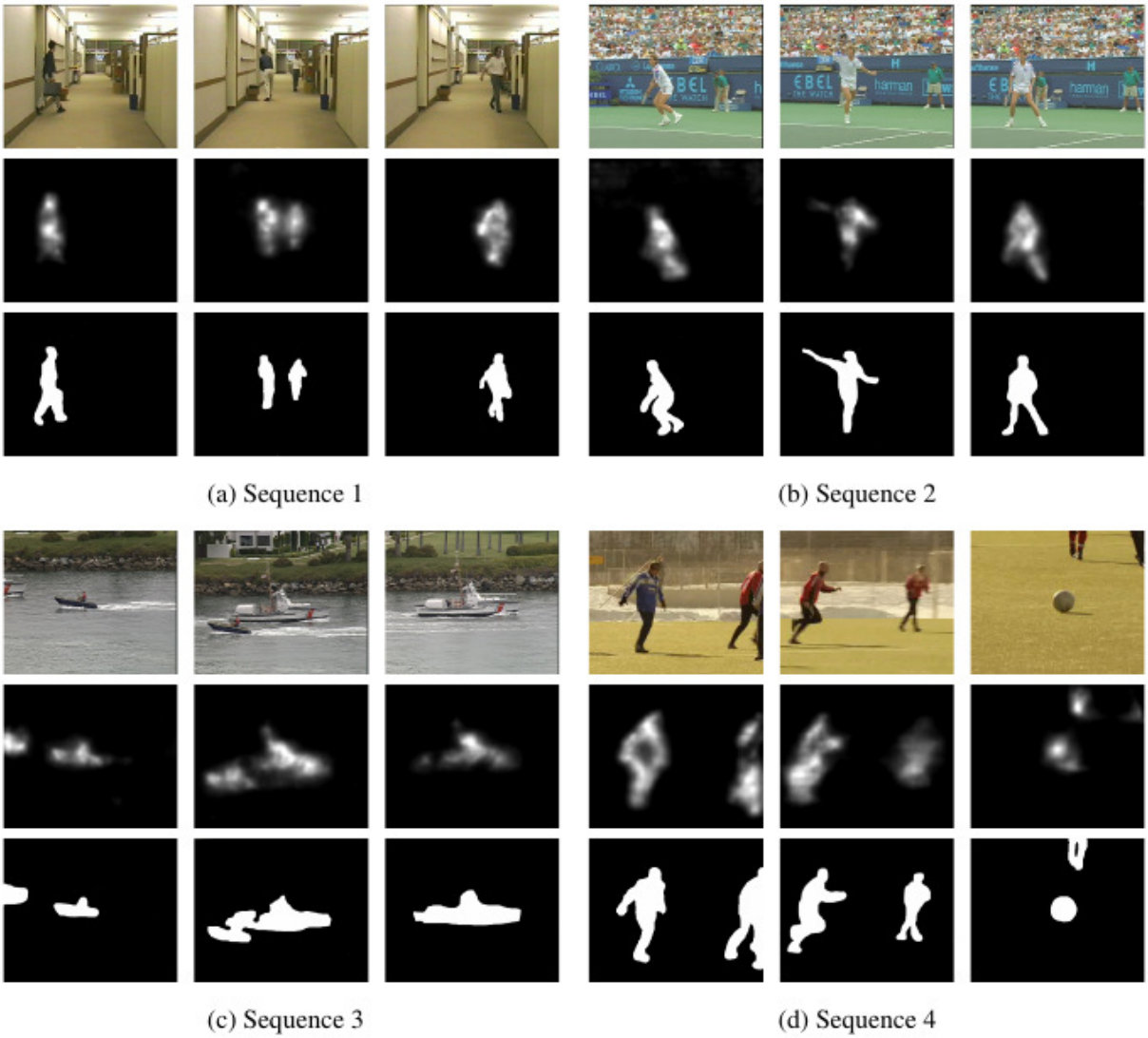
The proposed VA-based watermarking is agnostic to the watermark embedding methodology. Thus, it can be used on any existing watermarking algorithm. In our experiments, we use the non-blind embedding proposed by Xia *et al.*[53] and the blind algorithm proposed by Xie and Arce[54] as our reference algorithms.

A series of experimental results are generated for our video watermarking case study as described in Section 3, analysing 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[70] is provided, as common video attacks comprise 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 re-quantization 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.

An α_{max} and α_{min} approximating a PSNR of 35dB and 40dB, respectively, is utilised by applying Eq. (22) and Eq. (23). Four scenarios of varied watermark embedding strength are considered for the VA-based video watermarking evalution as follows:

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

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



(a) Sequence 1 (b) Sequence 2

(c) Sequence 3 (d) Sequence 4

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 http://svc.group.shef.ac.uk/VAvideo.html

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 viewer's overall opinion of the quality of experience (QoE) 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, dependant upon mathematical modelling, to determine visual quality. Such metrics include Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM),[71] Just Noticable Difference (JND)[72] and Video Quality Metric (VQM).[73] PSNR that calculates the average error between two images, is one of the most commonly used visual quality metrics and described in Eq. (20). Unlike PSNR, SSIM focuses on a quality assess-

Table 2: PSNR, SSIM and VQM average of 4 video sequences for non-blind watermarking.

	Low Strength	**Proposed**	**Average Strength**	**High Strength**
PSNR	40.15 ± 0.80	37.39 ± 0.87	37.47 ± 0.76	34.93 ± 0.73
SSIM	0.99 ± 0.00	0.97 ± 0.00	0.98 ± 0.00	0.95 ± 0.01
VQM	0.10	0.15	0.18	0.25

Table 3: PSNR, SSIM and VQM average of 4 video sequences for blind watermarking.

	Low Strength	**Proposed**	**Average Strength**	**High Strength**
PSNR	40.23 ± 1.03	36.80 ± 1.02	37.20 ± 0.92	34.85 ± 0.90
SSIM	0.99 ± 0.00	0.98 ± 0.00	0.98 ± 0.00	0.96 ± 0.01
VQM	0.08	0.13	0.15	0.22

ment 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 modelling. It incorporates numerous aspects of early visual processing, including both luma and chroma channels, a combination of temporal and spatial 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 Table 2 and Table 3 for non-blind 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 (α_{min}) $>$ VA-based algorithm / average strength $>$ high strength embedding (α_{max}). From the tables, PSNR improvements of approximately 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 decrease of 2% for the high strength watermarking model in most cases. In terms of 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 radiate 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 watermarking scheme on the overall perceived human viewing experience.

2) Subjective evaluation measures the visual quality by recording the opinion of human sub-

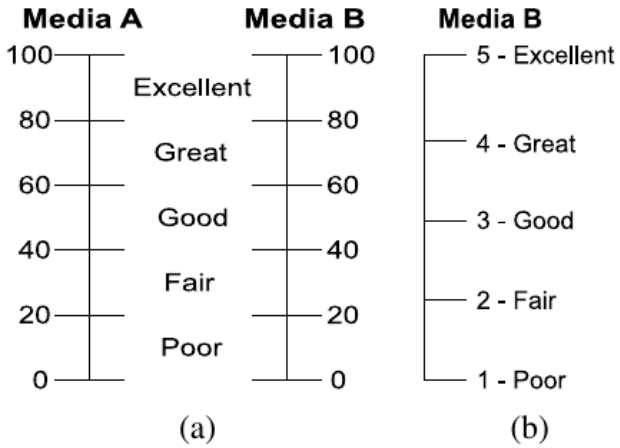


Fig 12: Subjective testing visual quality measurement scales (a) DCR continuous measurement scale (b) ACR ITU 5-point discrete quality scale.

jects on the perceived visual quality. The watermarked videos were viewed by these 30 subjects, following the standard International Telecommunication Union (ITU-T)[37] 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:

DSCQT: 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 randomised order, who 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 rating for the two test images.

DSIST: 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, firstly, 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 illustrates an overall timing diagram for each subjective testing procedure, showing the sequence of test image display 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 non-blind and blind watermarking schemes. The top and bottom rows in Fig 14 show subjective results for the non-blind 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 non-blind scenarios. Fig 14 shows the subjective test results for DCQST and DSIST averaged over 4 video test sequences. For the DSCQT, the lower the DCR, the better the visual quality, *i.e.*, less embedding distortions. In

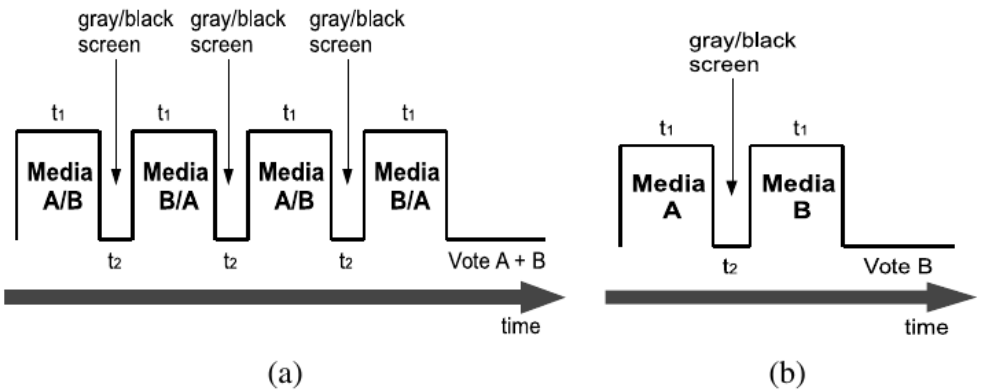


Fig 13: Stimulus timing diagram for (a) DCR method (b) ACR method.

the shown 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 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.*, less embedding visual distortions. DSIST plots for low-strength and VAM-based schemes show a similar ACR MOS approximately in the range 3-4, whereas the high strength watermark yields an ACR of less than 1 for non-blind and nearly 2 for blind watermarking. Compared with an average watermark strength, the proposed watermarking scheme shows an improved subjective image quality in all 4 graphs by around 0.5-1 units. As more data is 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 non-blind watermarking cases in both the objective and subjective visual quality evaluation, the low strength watermark and VAM-based watermarking sequences yield similar visual quality, where as the high strength embedded sequence appears severely more distorted. Low strength watermarking provides a high imperceptibility but is fragile as discussed in Section 4.2.2.

4.2.2 Robustness

Video reformatting and compression is a frequent and typically unintentional adversary attack, 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 Fig 16a) and Fig 16b) respectively. For simulating the watermark robustness, five constant Quantisation 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

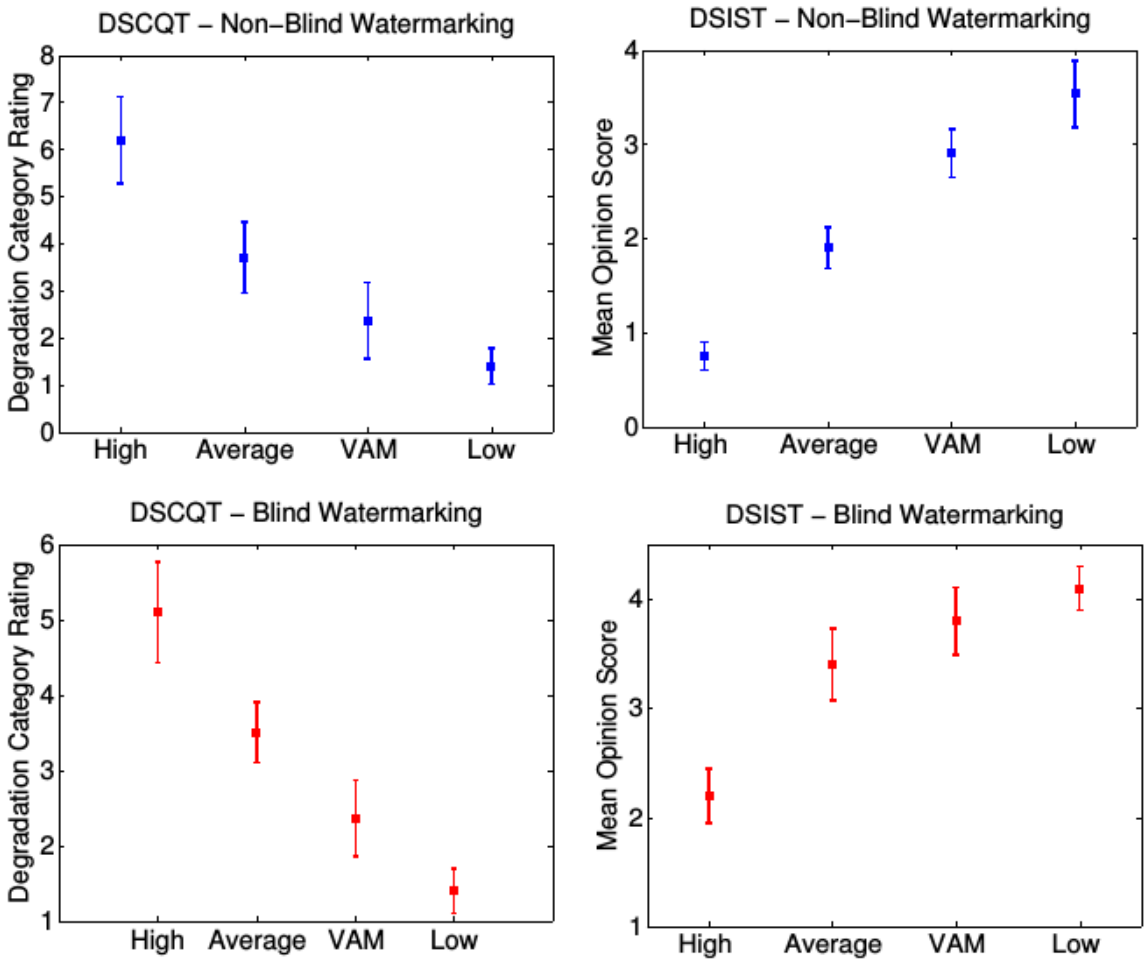


Fig 14: Subjective video watermarking embedding distortion measures for different embedding scenarios - high strength (High), average strength, VA-based strength selection (VAM) and low strength (Low) - *row 1:* non-blind watermarking, *row 2:* blind watermarking, *column 1:* DSCQT and *column 2:* DSIST.

lower Hamming distance indicates better robustness. From the plots in Fig 16, Hamming distance reductions up to 39% for the non-blind 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 towards H.264/AVC compression, as shown in Fig 16, when compared against a constant average strength watermark. It is of suitable note 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 high imperceptibility, verified by subjective and objective metrics, deem the VA-based methodology highly suitable towards providing an efficient

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.

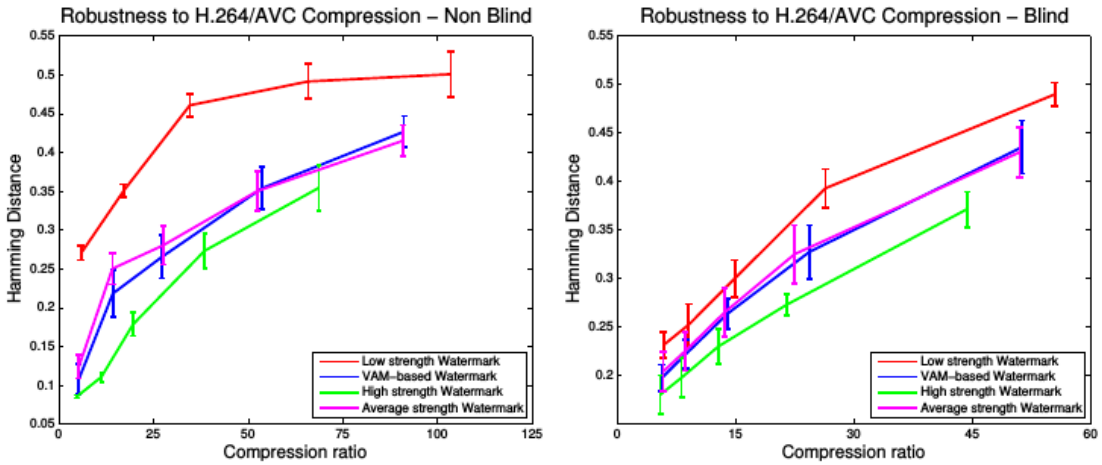


Fig 16: Robustness to H.264/AVC compression - average of 4 video sequences: a) non-blind watermarking and b) blind watermarking.

watermarking scheme.

5 Conclusions

In this paper, we have presented a novel video watermarking algorithm using a motion compensated visual attention model. The proposed method exploits both spatial and temporal cues for saliency modelled in a motion-compensated spatio-temporal wavelet multi resolution analysis framework. The spatial cues were modelled using the 2D wavelet coefficients. The temporal cues were modelled using the temporal wavelet coefficients by considering the global and local motion in the video. We have used of 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 is embedded into the host image according to the visual attentiveness of each region. By avoiding higher strength watermarking in 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 visual attention 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 movement are not fully accounted for and 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% (non-blind) and 22% (blind) improvement in robustness against H.264/AVC compression attacks, compared to the existing methodology that does not use the visual attention model. 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," *Cognitive Psychology* **12**(1), 97 – 136 (1980).

2 O. Hikosaka, S. Miyauchi, and S. Shimojo, "Orienting of spatial attention-its reflexive, compensatory, and voluntary mechanisms," *Cognitive Brain Research* **5**(1-2), 1–9 (1996).

3 L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience* **2**, 194–203 (2001).

4 R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual review of neuroscience* **18**(1), 193–222 (1995).

5 L. Itti and K. Christof, "Computational modelling of visual attention," *Natural Review Neuroscience* **2**, 194–203 (2001).

6 D. Ress, B. T. Backus, and D. J. Heeger, "Activity in primary visual cortex predicts performance in a visual detection task," *Nature neuroscience* **3**(9), 940–945 (2000).

7 J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?," *Natural Review Neuroscience* **5**(1), 1–7 (2004).

8 A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology* **12**(1), 97 – 136 (1980).

9 R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned Salient Region Detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1597 –1604 (2009).

10 A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analalysis Machine Intelligence* **35**, 185–207 (2013).
11 L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, 1254 –1259 (1998).
12 L. ZhiQiang, F. Tao, and H. Hong, "A saliency model based on wavelet transform and visual attention," *Information Sciences* **53**, 738–751 (2010).
13 S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2376–2383 (2010).
14 B. P. Ölveczky, S. A. Baccus, and M. Meister, "Segregation of object and background motion in the retina," *Nature* **423**(6938), 401–408 (2003).
15 F. Guraya and F. Cheikh, "Predictive visual saliency model for surveillance video," in *Proc. IEEE European signal processing conference (EUSPICO 2011)*, 554–558 (2011).
16 O. Meur, P. Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Research* **47**, 2483–2498 (2007).
17 L. Itti and N. Dhavale, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Proc. SPIE International Symposium on Optical Science and Technology*, 64–78 (2003).
18 Y. Tong, C. Faouzi, G. Fahad, K. Hubert, and T. Alain, "A spatiotemporal saliency model for video surveillance," *Cognitive Computation* **3**, 241–263 (2011).
19 C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing* **19**, 185–198 (2010).
20 S.-h. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren, "Video saliency detection via dynamic consistent spatio-temporal attention modelling.," in *AAAI*, (2013).
21 Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Transactions on Image Processing* **23**, 3910–3921 (2014).
22 Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM International Conference on Multimedia*, 815–824 (2006).
23 M. Dorr, E. Vig, and E. Barth, "Colour saliency on video," in *Bio-Inspired Models of Network, Information, and Computing Systems*, **87**, 601–606 (2012).
24 C. C. Loy, X. Tao, and G. S. Gong, "Salient motion detection in crowded scenes," in *International Symposium on Communications Control and Signal Processing, (ISCCSP)*, 1–4 (2012).
25 D. Que, L. Zhang, L. Lu, and L. Shi, "A ROI image watermarking algorithm based on lifting wavelet transform," in *Proc. International Conference on Signal Processing*, **4**, 16–20 (2006).
26 R. Ni and Q. Ruan, "Region of interest watermarking based on fractal dimension," in *Proc. International Conference on Pattern Recognition*, 934–937 (2006).
27 R. Wang, Q. Cheng, and T. Huang, "Identify regions of interest (ROI) for video watermark embedment with principle component analysis," in *Proc. ACM International Conference on Multimedia*, 459–461 (2000).
28 C. Yiping, Z. Yin, Z. Sanyuan, and Y. Xiuzi, "Region of interest fragile watermarking for image authentication," in *International Multi-Symposiums on Computer and Computational Sciences (IMSCCS)*, **1**, 726–731 (2006).

29 L. Tian, N. Zheng, J. Xue, C. Li, and X. Wang, "An integrated visual saliency-based watermarking approach for synchronous image authentication and copyright protection," *Image Communication* **26**, 427–437 (2011).

30 B. Ma, C. L. Li, Y. H. Wang, and X. Bai, "Salient region detection for biometric watermarking," *Computer Vision for Multimedia Applications: Methods and Solutions* , 218 (2011).

31 A. Sur, S. Sagar, R. Pal, P. Mitra, and J. Mukhopadhyay, "A new image watermarking scheme using saliency based visual attention model," in *India Conference (INDICON), 2009 Annual IEEE*, 1–4 (2009).

32 M. Oakes, D. Bhowmik, and C. Abhayaratne, "Visual attention-based watermarking," in *IEEE International Symposium on Circuits and Systems (ISCAS),*, 2653–2656 (2011).

33 D. Bhowmik, M. Oakes, and C. Abhayaratne, "Visual attention-based image watermarking," *IEEE Access* **PP**(99), 1–1 (2016).

34 M. Oakes, *Attention Driven Solutions for Robust Digital Watermarking Within Media*. PhD thesis, University of Sheffield (2014).

35 L. Tian, N. Zheng, J. Xue, and C. Li, "Authentication and copyright protection watermarking scheme for h. 264 based on visual saliency and secret sharing," *Multimedia Tools and Applications* **74**(9), 2991–3011 (2015).

36 A. Barari and S. V. Dhavale, "Video saliency detection for visual cryptography-based watermarking," *Innovative Research in Attention Modeling and Computer Vision Applications* , 132 (2015).

37 H. G. Koumaras, "Subjective video quality assessment methods for multimedia applications," Tech. Rep. ITU-R BT.500-11, Geneva, Switzerland (2008).

38 M. Oakes and C. Abhayaratne, "Visual saliency estimation for video," in *Proc. 13th International Workshop on Image Analysis for Multimedia Interactive Services*, 1–4 (2012).

39 H. Wilson, "Psychophysical models of spatial vision and hyper-acuity," *Spatial Vision* **10**, 64–81 (1991).

40 R. Jin, Y. Qi, and A. Hauptmann, "A probabilistic model for camera zoom detection," in *Proc. International Conference on Pattern Recognition*, **3**, 859–862 (2002).

41 H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *In ECCV*, 404–417 (2006).

42 S. Choi and W. W. J, "Motion-compensated 3-d subband coding of video," *IEEE Transactions on Image Processing* **8**(2), 155–167 (1999).

43 Y. Andreopoulos, A. Munteanu, J. Barbarien, M. Schaar, J. Cornelis, and P. Schelkens, "In-band motion compensated temporal filtering," *Signal Processing: Image Communication* **19**(7), 653–673 (2004).

44 D. Bhowmik and C. Abhayaratne, "2D+t wavelet domain video watermarking," *Advances in Multimedia* **2012**, 6 (2012).

45 D. Mahapatra, S. Winkler, and S. Yen, "Motion saliency outweighs other low-level features while watching videos," in *Proc. SPIE conference on Human Vision and Electronic Imaging*, **6806**, 1–10 (2008).

46 D. S. Taubman and M. W. Marcellin, *JPEG2000 Image Compression Fundamentals, Standards and Practice*, Springer, USA (2002).

47 P. Chen and J. W. Woods, "Bidirectional MC-EZBC with lifting implementation," *Circuits and Systems for Video Technology, IEEE Transactions on* **14**(10), 1183–1194 (2004).

48 G. Bhatnagar and Q.M. J. Wuand B. Raman, "Robust gray-scale logo watermarking in wavelet domain," *Computers & Electrical Engineering* (2012).

49 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*, 79–90 (2005).

50 D. Bhowmik and C. Abhayaratne, "A generalised model for distortion performance analysis of wavelet based watermarking," in *Proc. Int'l Workshop on Digital Watermarking (IWDW '08), Lect. Notes in Comp. Sci. (LNCS)*, **5450**, 363–378 (2008).

51 C. Abhayaratne and D. Bhowmik, "Scalable watermark extraction for real-time authentication of JPEG2000 images," *Journal of Real-Time Image Processing* **6**(4), 19 pages (2011).

52 D. Bhowmik and C. Abhayaratne, "Quality scalability aware watermarking for visual content," *IEEE Transactions on Image Processing* **25**(11), 5158–5172 (2016).

53 X. Xia, C. G. Boncelet, and G. R. Arce, "Wavelet transform based watermark for digital images," *Optic Express* **3**, 497–511 (1998).

54 L. Xie and G. R. Arce, "Joint wavelet compression and authentication watermarking," in *Proc. IEEE ICIP*, **2**, 427–431 (1998).

55 M. Barni, F. Bartolini, and A. Piva, "Improved wavelet-based watermarking through pixel-wise masking," *IEEE Trans. Image Processing* **10**, 783–791 (2001).

56 C. Jin and J. Peng, "A robust wavelet-based blind digital watermarking algorithm," *International Journal of Information Technology* **5**(2), 358–363 (2006).

57 R. S. Shekhawat, V. S. Rao, and V. K. Srivastava, "A biorthogonal wavelet transform based robust watermarking scheme," in *Proc. IEEE Conference on Electrical, Electronics and Computer Science (SCEECS)*, 1–4 (2012).

58 D. Bhowmik and C. Abhayaratne, "On Robustness Against JPEG2000: A Performance Evaluation of Wavelet-Based Watermarking Techniques," *Multimedia Syst.* **20**(2), 239–252 (2014).

59 Q. Gong and H. Shen, "Toward blind logo watermarking in JPEG-compressed images," in *Proc. Int'l Conf. on Parallel and Distributed Comp., Appl. and Tech., (PDCAT)*, 1058–1062 (2005).

60 J. R. Kim and Y. S. Moon, "A robust wavelet-based digital watermarking using level-adaptive thresholding," in *Proc. IEEE ICIP*, **2**, 226–230 (1999).

61 V. Saxena, M. Gupta, and D. T. Gupta, "A wavelet-based watermarking scheme for color images," *The IUP Journal of Telecommunications* **5**, 56–66 (2013).

62 C. Jin and J. Peng, "A robust wavelet-based blind digital watermarking algorithm," *Information Technology Journal* **5**(2), 358–363 (2006).

63 F. Huo and X. Gao, "A wavelet based image watermarking scheme," in *Proc. IEEE ICIP*, 2573–2576 (2006).

64 D. Kundur and D. Hatzinakos, "Digital watermarking using multiresolution wavelet decomposition," in *Proc. IEEE ICASSP*, **5**, 2969–2972 (1998).

65 V. S. Verma and J. R. Kumar, "Improved watermarking technique based on significant difference of lifting wavelet coefficients," *Signal, Image and Video Processing* , 1–8 (2014).

66 P. Meerwald, "Quantization watermarking in the JPEG2000 coding pipeline," in *Proc. Int'l Working Conf. on Comms. and Multimedia Security*, 69–79 (2001).
67 D. Aggarwal and K. S. Dhindsa, "Effect of embedding watermark on compression of the digital images," *Computing Research Repository* **1002** (2010).
68 K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *Proc. IEEE International Conference on Multimedia and Expo*, 638–641 (2009).
69 X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in neural information processing systems*, 681–688 (2008).
70 H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circ. and Syst. for Video Tech* **17**(9), 1103–1120 (2007).
71 Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing* **13**, 600–612 (2004).
72 A. B. Watson, "Visual optimization of dct quantization matrices for individual images," in *American Institute of Aeronautics and Astronautics (AIAA)*, **9**, 286291 (1993).
73 A. B. Watson, J. Hu, and J. F. M. Iii, "DVQ: A digital video quality metric based on human vision," *Journal of Electronic Imaging* **10**, 20–29 (2001).

Matthew Oakes graduated with the M.Eng in Electronic and Electrical Engineering from the University of Sheffield in 2009. In 2014 he obtained the PhD in Electronic and Electrical Engineering also at the University of Sheffield. Matthew is currently working at the University of Buckingham as a Knowledge Transfer Partnership Associate in joint project between the Applied Computing Department at Buckingham and Deepnet Security Ltd in Bletchley Park. 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 the 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, the University of Sheffield, UK and a system engineer in ABB Ltd., India. Currently he is working as a lecturer at Sheffield Hallam University, Sheffield, UK. He has been awarded a UK Engineering and Physical Science Research council (EPSRC) Dorothy Hodgkin Postgraduate Award for his PhD study and a Digital Catapult EPSRC fellowship to conduct watermarking research. His current research interests include computer vision, machine learning, embedded imaging hardware on FPGA and multimedia security.

Charith Abhayaratne received the B.E, in Electrical and Electronic Engineering from the University of Adelaide in Australia in 1998 and the Ph.D. in Electronic and Electrical Engineering from the University of Bath in the UK in 2002. Currently, he is a lecturer within the Department of Electronic and Electrical Engineering at the University of Sheffield in the UK. He was an European Research Consortium for Informatics and Mathematics (ERCIM) postdoctoral fellow in 2002-2004 at the Centre for mathematics and computer science (CWI) in The Netherlands and at the National Research Institute for computer science and control (INRIA) in Sophia Antipolis in

France. His research interests and expertise include video and image compression, watermarking, image and video analysis, wavelets and signal transforms. He has contributed to MPEG on scalable video coding standardization and wavelet video exploration in 2004-2006. He has been a reviewer for leading IEEE/IET/ACM/SPIE journals and conferences, UK Engineering and Physical Science Research council (EPSRC), Hong Kong research council and book publishers. He has been involved in technical programme committees for several leading international conferences and served as session chairs.

List of Figures

List of Tables