Signal Processing: *Image Communication* 38 (2015) 15–31

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

A perceptually based spatio-temporal computational framework for visual saliency estimation

Petros Koutras *, Petros Maragos

School of Electrical and Computer Engineering, National Technical University of Athens, Zografou Campus, Athens 15773, Greece

ARTICLE INFO

Available online 20 August 2015

Keywords:
Spatio-temporal visual frontend
3D Gabor filters
LAB color space
Visual saliency
Eye-tracking database

ABSTRACT

The purpose of this paper is to demonstrate a perceptually based spatio-temporal computational framework for visual saliency estimation. We have developed a new spatio-temporal visual frontend based on biologically inspired 3D Gabor filters, which is applied on both the luminance and the color streams and produces spatio-temporal energy maps. These volumes are fused for computing a single saliency map and can detect spatio-temporal phenomena that static saliency models cannot find. We also provide a new movie database with eye-tracking annotation. We have evaluated our spatio-temporal saliency model on the widely used CRCNS-ORIG database as well as our new database using different fusion schemes and feature sets. The proposed spatio-temporal computational framework incorporates many ideas based on psychological evidences and yields significant improvements on spatio-temporal saliency estimation.

1. Introduction

In biological vision systems there exist significant neurobiological and psychophysical evidences that the first stages of visual information processing include many feature detection processes. Since various stages of biological vision systems involve spatio-temporal processing and nature has a tendency to represent information in optimal ways, efficient perception-inspired spatio-temporal processing as well as easily computable features that can compactly represent salient structure in moving images should be one of the important early goals of video processing.

Visual attention is a cognitive mechanism employed by humans, animals and artificial systems for selecting the most important part of information from a visual stimulus and then perform more complex and demanding processes. This field has been for years an active research subject for psychophysics and cognitive scientists, because attention mechanisms play a dominant role in human visual system.

Attention may have two modes, a top-down expectation-driven, and a bottom-up stimulus-driven, and so there is often a confusion between attention and visual saliency. Visual attention is a wider concept, which often includes many topics, such as top-down cognitive information processing, memory, object searching, task demands or expectations. On the other hand, visual saliency is a bottom-up process and is based on the sensory cues of a stimulus that make certain image or video regions more conspicuous. In addition to its cognitive and biological nature, several computational frameworks have also been proposed for modeling visual saliency [1], because it plays a significant role in many computer vision applications, such as object and action recognition [2–5] and movie summarization [6–8].

We propose a spatio-temporal computational frontend for visual saliency, which is suitable for estimating spatio-temporal events in video streams. Its design is built upon

* Corresponding author.
E-mail addresses: pkoutras@cs.ntua.gr (P. Koutras), maragos@cs.ntua.gr (P. Maragos).

http://dx.doi.org/10.1016/j.image.2015.08.004
0923-5965/

many ideas from biological and perceptual image processing, related to human vision modeling. During the past decades several computational approaches have been developed for visual saliency estimation in the spatial domain, which had incorporated many advanced techniques for processing the luminance and color modalities. More recently there appeared models for spatio-temporal estimation in video stimuli that are mainly based on simple motion estimation or spatio-temporal filtering rather than using only the classic static cues (intensity, color, orientation). Our approach is designed for spatio-temporal estimation and incorporates advances in both static and spatio-temporal pathways. A brief summary of biologically inspired feature detection methods as well as the spatial and spatio-temporal computational models is given in Section 2.

Our framework exhibits unification and computational economy in at least three important ways: it produces both spatio-temporal and static energy volumes by using the same multi-scale filterbank based on quadrature Gabor filters in three dimensions (space and time). In addition, the same framework can be applied for two different modalities, i.e. the image luminance and color stream modalities, producing independent spatio-temporal energy volumes. For the color stream we have incorporated many modern ideas such as LAB color space or PCA analysis. Further, our spatio-temporal framework can provide motion information in different scales and directions without having to process it as a separate cue or use a small number of frames like other video saliency approaches require. In this way, our approach achieves to detect both the fastest changes in the video stimuli (e.g. flicker) and the slowest motion changes related to action events. The produced energy maps can be integrated into a single spatio-temporal saliency map, by using different energy mixtures and fusion schemes. The complete model and the filtering details are analyzed in Section 3.

Our computational approach is evaluated in two different ways. At first we employed simple spatio-temporal stimuli, where our method manages to detect time-varying events that static saliency method cannot find. The second application is the prediction of human eye fixations while the subjects watch video stimuli, using a single spatio-temporal saliency map. We use two databases with eye-tracking data annotation: the CRCNS-ORIG [9] and our newly created *Eye-Tracking Movie Database (ETMD)*. The latter was collected for the purposes of the presented study and comprises short video clips from Hollywood movies along with eye tracker data for 10 subjects. In Section 4 we describe the evaluation procedure and our experimental results in both these databases. In general, our method for spatio-temporal saliency estimation is quite promising as it achieves higher performance than many other state-of-the-art saliency models.

2. Background/related work

Assuming that visual information processing by several classes of optical neurons can be modeled by linear operators, there was a hot debate in the perceptual and neurophysiological research community during the 1960s and 1970s as to whether the early stages of visual information processing in primates can be modeled as spatial local feature detectors or as filterbanks in the frequency domain. From the side of spatial processing, Hubel and Wiesel [10,11] found in cat's and monkey's visual cortex simple cells whose behavior they described as approximately linear *feature detectors* with line-, edge- or bar-shaped receptive fields that exhibited scale and orientation selectivity. Since then these results have been confirmed and refined by many other researchers [12–14]. From the side of frequency domain, several researchers have argued, based on psychophysical experiments, that the early visual system can be approximately modeled using Fourier analysis ideas [15–17], mainly in one-dimension (1D) until Daugman [18] proposed a two-dimensional (2D) spatial filtering and Fourier analysis. In another experimental direction, Pollen and Ronner [19] found that adjacent simple cells in the visual cortex are tuned to the same spatial frequency and orientation, but their responses are in *quadrature*.

Daugman [20] also observed that, from a mathematical viewpoint, the antagonism between the spatial and frequency domain interpretations of visual information processing is illusionary, since neurons in the retina or visual cortex can both resemble filterbanks of bandpass filters, or, equivalently, convolutions with neuron responses that have excitatory or inhibitory regions in their center-surround receptive fields. He further extended the existing 1D Gabor theory [21,22] and proposed the *2D oriented Gabor filters* as optimal models for simple cell impulse responses, where 'optimality' here means having minimal *space–frequency uncertainty*. Since then, Gabor filters in quadrature pairs have been extensively used in many early computer vision tasks, e.g. in 2D spatial texture analysis [23] and in spatio-temporal models for motion [24] and optical flow estimation [25,26].

For the modeling of receptive fields (RFs) of cells in the visual system in parallel to the use of Gabor filters, a few other approaches were also proposed such as *Difference of Gaussians* (*DOG*) filters by Wilson and Bergen [27] and the *Derivatives of Gaussians (GD)* (or in discrete form *Difference of Offsets of Gaussians* (*DOOG*)) by Young [28,29]. The former has limited applicability and was used mainly for isotropic center-surround RFs and edge detection [30]. The latter found a wider acceptance and was used for modeling the RFs of simple cells in primate visual systems, by applying *Gaussian Derivatives* up to tenth order instead of Gabor filters. As Koenderink and van Doorn [31] proved, for high orders the Gaussian derivatives become approximate Gabor filters. Later, the GD model was extended to spatio-temporal vision [32].

In addition, all the above filter models come at *multiple scales* (corresponding to the various frequency channels) and may be either isotropic (e.g. in the retina) or *oriented* (e.g. in the visual cortex). Moreover, there have also been other perception-inspired models for feature detection that are nonlinear and based on ideas of phase congruency and quadrature energy, as in Morrone et al. [33,34]. Filterbanks with 2D spatial filters in quadrature pairs of the Gabor, GD, or similar type followed by nonlinear operations like energy computation or

half-wave rectification have been widely used in texture analysis [35,23,36] and boundary detection [37].

In parallel with the research in feature detection models, Treisman and Gelade [38] pointed out which visual features are important and how they can be integrated in human visual attention. Later, Koch and Ullman [39], based on these features, proposed a visual saliency model and introduced the *spatial saliency map*, which describes which image regions are more conspicuous. These two theories became the basis of many cognition-inspired attention models [40–43], while Itti et al. [44] provided an implementation of a bottom-up computational model for spatial visual saliency using three feature channels: intensity, color, and orientation. This model was later extended into a spatio-temporal model for visual saliency estimation in video streams by the use of two additional features: motion and flicker [45]. These computational models have found a wide acceptance and many other cognition-related approaches aiming at either spatial [2,46–48] or spatio-temporal analysis [49–51] have been based on them. In the same biological-inspired concept, the adaptive whitening saliency model [52] used spatial 2D log-Gabor filters and decorrelated the multiscale filter responses using PCA analysis in order to obtain a final saliency map in images.

Although the early methods for visual saliency drew inspiration from biological models of the human vision system, later approaches for saliency estimation were based on a Bayesian framework [53] and took advantage of the feature statistics of the images, such as *Bayesian surprise* over space or time [9,54] and salient object detection [55]. Zhang et al. [56] proposed a general framework for saliency using natural (SUN) scene statistics, while later they extended their model by including spatio-temporal features [57]. Several *information theoretic* measures have also been used like *entropy* of local features distributions [3]. Bruce and Tsotsos [58] proposed a model based on *self-information* from a prior local model for calculating saliency in image regions and later extended their model to a spatio-temporal version [59]. Hou and Zhang [60] proposed the incremental coding length as a measurement of the perspective entropy gain of each visual feature in order to achieve attention selectivity in both static and dynamic stimuli. Gao and Vasconcelos [61,62] computed visual saliency based on the *mutual information* between features and image regions. They used DOG and Gabor filters to measure the discriminative power of features in center-surround image regions. Later their initial model was combined with motion information to estimate spatio-temporal saliency in dynamic scenes [63]. Seo and Milanfar [64] used local regression kernels, in images or videos, and matrix cosine similarity to measure the self-resemblance of an image region with its local surroundings. Riche et al. [65] employed Gabor filtering in a PCA transformed color space and computed visual saliency as the rare regions of the image using both local and global contrast.

In another class of approaches, saliency is estimated in the frequency domain by frequency- or phase-selective tuning of the saliency map [66–68]. Such models are based on Fourier or discrete cosine transforms [66,69] while the quaternion Fourier transform is also used for combining color, intensity and motion features [70,71,68]. Fourier spectrum in spatio-temporal domain is also applied on video slices along x–t and y–t planes to separate foreground motion objects from backgrounds [72].

Most saliency methods have been developed mainly for still images. At present, there has not been much work on spatio-temporal saliency models. Instead, some static methods have been extended to a spatio-temporal version by using additional features related to temporal information or motion. For example, in [45,9,48] differences between the spatial orientation maps are employed as temporal features for saliency detection in videos. In [64] the authors extended their self-resemblance method by employing 3D local steering kernels for action and saliency detection in videos. In [73] a spatio-temporal filtering using temporal weighted sum is proposed for abnormal motion selection in crowed scenes, while [74] combine camera motion information with static features to study the differences between static and dynamic saliency in videos.

Recently, machine learning techniques have been adopted to detect saliency for still images [75] or videos [76–78] employing both low-level features, such as orientation, color, intensity and optical flow motion and high level features, such as face and object recognition. Moreover, probabilistic learning techniques based on bottom-up saliency and gist descriptors are also employed for task-specific [79] or multi-task [80] eye-tracking prediction in spatio-temporal stimuli. Further reviews of additional approaches for visual saliency estimation can be found in [81,82,1].

3. Spatio-temporal visual frontend

Our energy-based model for spatio-temporal visual saliency estimation is more relevant to the cognition-inspired

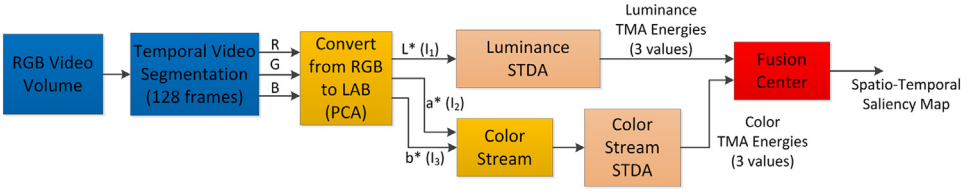

Fig. 1. Overall process for spatio-temporal saliency estimation. First the original video is cut into small temporal segments and the RGB color space is transformed into LAB space or a PCA transformed color space. Then follows the Spatio-Temporal Dominant Analysis (STDA), which is applied both on luminance and color stream channels. The resulting energy volumes are combined under different fusion schemes to form a single spatio-temporal saliency map.

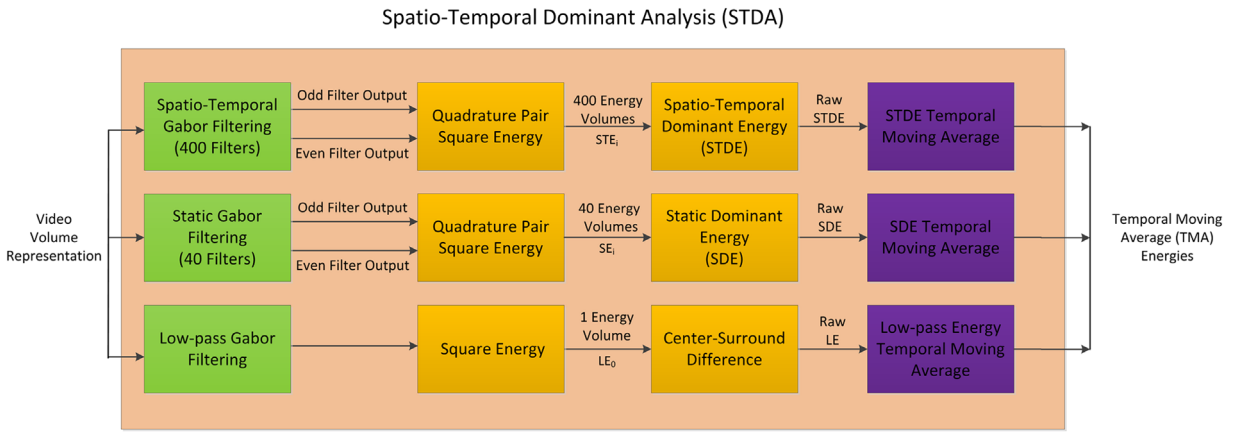

Fig. 2. Spatio-Temporal Dominant Analysis (STDA) which contains three individual stages: spatio-temporal Gabor filtering, quadrature pair energy computation and dominant energy selection and the Temporal Moving Average (TMA) applied on the raw energies.

saliency methods, based on Koch & Ullman theory. It uses biologically plausible spatio-temporal filters, like oriented 3D Gabor filters, in order to extract visual features which are composed into a single saliency map. The overall process is shown in Fig. 1. In a first phase the initial RGB video volume is transformed into the LAB space or into a PCA transformed space and split into two streams: luminance and color stream. Then follows the main process step, called *Spatio-Temporal Dominant Analysis* (*STDA*), which is applied both on luminance and color stream channels. (Recall that a separate motion cue is not needed since relevant information is indirectly provided by our spatio-temporal processing.) The last stage includes the fusion process in which different fusion methods and features mixtures are employed in order to produce a single spatio-temporal saliency map.

3.1. Preprocessing and color modeling

Let the 3-value vector $\mathbf{I}_{RGB}(x, y, t)$ be the RGB volume representation of the whole initial video, where x, y are frame-based spatial coordinates and t is the time index corresponding to each frame. The first preprocessing step is to cut the original video volume into successive segments, in order to avoid memory overloads during the spatio-temporal Gabor filtering. Each video segment consists of 128 frames, which correspond to 4–5 s duration for typical video rates. This is sufficient for our temporal analysis since the lowest temporal frequencies used cannot model more slowly changing events.

Then, follows the color modeling where we use a color space in which luminance and chromaticity components can be well separated, instead of the RGB color space in which the three color components (R, G, B) are highly correlated. Specifically in our first approach we choose the CIE-LAB (L^*, a^*, b^*) color space because this space, compared with other color spaces like HSI or YCbCr, has the additional property to be perceptually uniform. The CIE-LAB space is created from a nonlinear transformation on CIE-XYZ color space [83]. In our second approach for the color modeling we use the principal component analysis (PCA), which is inspired by neurophysiological evidences about the neural responses. First, we apply PCA in the initial *RGB* color vector. Then we divide each component with the root of its eigenvalue in order to have a decorrelated and whitened representation $I_i(x, y, t)$, $i = 1, 2, 3$ of the color space [52].

In the resulting video volume $\mathbf{I}_{LAB}(x, y, t)$ or $I_i(x, y, t)$ the first component (L^* or I_1) expresses the perceptual response to luminance, while a^*, b^* (or I_2, I_3) describe the color information. In order to describe the color stream in videos by a single measure with positive values regardless to the specific color, we use the L2 norm of the components I_1, I_2 that are related with color

$$C_{PCA}(x, y, t) = \sqrt{I_1^2 + I_2^2} \tag{1}$$

In the case of LAB color space we employed an approach that models the double color opponent cells that exist in primary visual cortex V1 and has been used in color constancy applications [84]. Instead of using the R, G, B components, we use the chromaticity components (a^*, b^*) that indirectly include the $R - G$ and $B - Y$ differences. The responses of the double opponent cells DO_a, DO_b can be computed as a weighed sum of single opponent cells' responses SO_a, SO_b with different scales

$$DO_a(x, y) = SO_a(x, y, \sigma_1) - w \cdot SO_a(x, y, \sigma_2) \tag{2}$$

$$DO_b(x, y) = SO_b(x, y, \sigma_1) - w \cdot SO_b(x, y, \sigma_2) \tag{3}$$

where σ_1, σ_2 are the scales of the center and surround receptive fields (RF) respectively and $w \in [0, 1]$ controls the contribution of the surround cell. Cells with $w=1$ respond only to color contrast while cells that have small w values enhance more the color regions of an image. We select $w=0.6$ in order to have both edges and regions in the color stream. We have assumed $\sigma_2 = 3\sigma_1$ since it is found by neurophysiological experiments that the surround receptive field is about 3 times larger than the center RF [85]. The responses of the single opponent cells can be implemented as a 2D Gaussian filtering

$$SO_a(x, y, \sigma) = a^* * G_{RF}(x, y, \sigma),$$
$$SO_b(x, y, \sigma) = b^* * G_{RF}(x, y, \sigma) \tag{4}$$

where G_{RF} is a 2D Gaussian function with standard deviation σ. The resulting color stream that expresses both the color intensity and the contrast is giving by

$$C_{ab}(x, y, t) = \sqrt{DO_a^2(x, y, t) + DO_b^2(x, y, t)} \quad (5)$$

Finally, for each video segment $\mathbf{I}_{LAB}(x, y, t)$ or $I_i(x, y, t)$ we keep the first component as the luminance stream and color stream $C_{ab}(x, y, t)$ ($C_{PCA}(x, y, t)$). These two visual information streams are forwarded to the main stage of our frontend, the STD analysis, for spatio-temporal energy feature extraction.

3.2. Spatio-Temporal Dominant Analysis

In this subsection we will describe the core stage of our perception-inspired frontend for visual saliency. As shown in Fig. 2, STDA can be divided into tree individual stages. The first stage consists of the spatio-temporal Gabor filtering, while the others include postprocessing procedures like quadrature pair energy computation and dominant energy selection followed by an optional Temporal Moving Average (TMA) applied on the resulting raw energies. We note that the same STDA method is applied without changes to both luminance and color stream modalities.

3.2.1. 3D Gabor filtering

The first step of STDA is the filtering process of the video volume. Among the filtering approaches that have been proposed based on psychophysical experiments, the two with the wider acceptance are the Gabor filters and the Gaussian Derivatives (GD). We choose to use oriented Gabor filters in a spatio-temporal version, due to their biological plausibility and their uncertainty-based optimality [21,20]. In addition, for high order derivatives the GD filters are approximations of the Gabor filters [31].

GD filters combined with their Hilbert transform (quadrature pair) are widely used in many spatial and spatio-temporal tasks [37,86], mainly because they can be implemented in an efficient way since they are steerable [87]. Gabor filters, or the other hand, are not strictly mathematically steerable but as Heeger [25,26] showed they can become separable, which means that a high dimensional Gabor filter can be built from 1D Gabor impulses responses.

In addition, 3D spatio-temporal filtering is also applied in [64] by employing a 3D extension of the Local Steering Kernel (LSK) [88]. These kernels are nonlinear and describe temporal voxel differences based on spatial and temporal gradient information inside a local 3D neighborhood. However, their nonlinear nature requires a lot of processing time and memory storage for the feature extraction. Thus, a small space–time neighborhood (e.g. $3 \times 3 \times 3$) is usually employed and the video resolution is down-sampled to a single and very coarse spatial scale. In [73] a low-pass spatial filter with a weighted temporal average is employed for the spatio-temporal filtering of motion features extracted with optical flow.

On the other hand, our proposed 3D filterbank is based on linear and biologically inspired Gabor filters which can describe both the gradient and texture information of the image in multiple scales as well as spatio-temporal changes and patterns related with motion and action. Our temporal analysis is also multi-scaled as we have included Gabor filters at different scales and spatio-temporal directions and does not use only a small number of successive frames. With this approach we can detect both fast changes in the video stimuli (e.g. flicker) and slow and complex motion changes related with action events. Spatio-temporal Gabor filtering has also been employed in [89] but separately in the $x-t$ and $y-t$ planes and not in a 3D multi-scale and multi-orientation manner; further, they test their model on synthetic stimuli and simple videos with elementary motion examples.

So, we apply quadrature pairs of 3D (spatio-temporal) Gabor filters with identical central frequencies and bandwidth. These filters can arise from 1D Gabor filters [21] in a similar way as Daugman proposed 2D oriented Gabor filters [18]. An 1D complex Gabor filter consists of a complex sine wave modulated by a Gaussian window. Its impulse response with unity norm has the following form:

$$g(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2} \right) \exp(j\omega_{t_0} t) = g_c(t) + j g_s(t) \quad (6)$$

The above complex filter can be split into one odd(sin)-phase ($g_s(t)$) and one even(cos)-phase ($g_c(t)$) filters, which form a quadrature pair filter. Almost all Gabor filters are bandpass filters whose center frequency coincides with their modulating frequency ω_{t_0}; the only exception where they become lowpass filters is when $\omega_{t_0} = 0$ which makes them Gaussians. Thus, we can cover the whole spatio-temporal 3D spectral domain with Gabor filters whose frequency responses are centered around specific frequencies, including the zero spatial and temporal frequencies which correspond to the case when we have no variation in this direction (static information).

The 3D Gabor extension (as for example used for optical flow in [26]) yields an *even* (cos) 3D Gabor filter whose impulse response is

$$g_c(x, y, t) = \frac{1}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_t} \exp\left[-\left(\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2} + \frac{t^2}{2\sigma_t^2} \right) \right]$$

$$\cdot \cos(\omega_{x_0} x + \omega_{y_0} y + \omega_{t_0} t) \quad (7)$$

where $\omega_{x_0}, \omega_{y_0}, \omega_{t_0}$ are the spatial and temporal angular center frequencies and $\sigma_x, \sigma_y, \sigma_t$ are the standard deviations of the 3D Gaussian envelope. Similarly for the impulse response of *odd (sin)* filter which we denote by $g_s(x, y, t)$. The frequency response of the even (cos) 3D Gabor Filter will have the following form:

$$\begin{aligned} G_c(\omega_x, \omega_y, \omega_t) = {} & \tfrac{1}{2} \exp[-(\sigma_x^2(\omega_x - \omega_{x_0})^2/2 \\ & + \sigma_y^2(\omega_y - \omega_{y_0})^2/2 + \sigma_t^2(\omega_t - \omega_{t_0})^2/2)] \\ & + \tfrac{1}{2} \exp[-(\sigma_x^2(\omega_x + \omega_{x_0})^2/2 \\ & + \sigma_y^2(\omega_y + \omega_{y_0})^2/2 + \sigma_t^2(\omega_t + \omega_{t_0})^2/2)] \end{aligned} \quad (8)$$

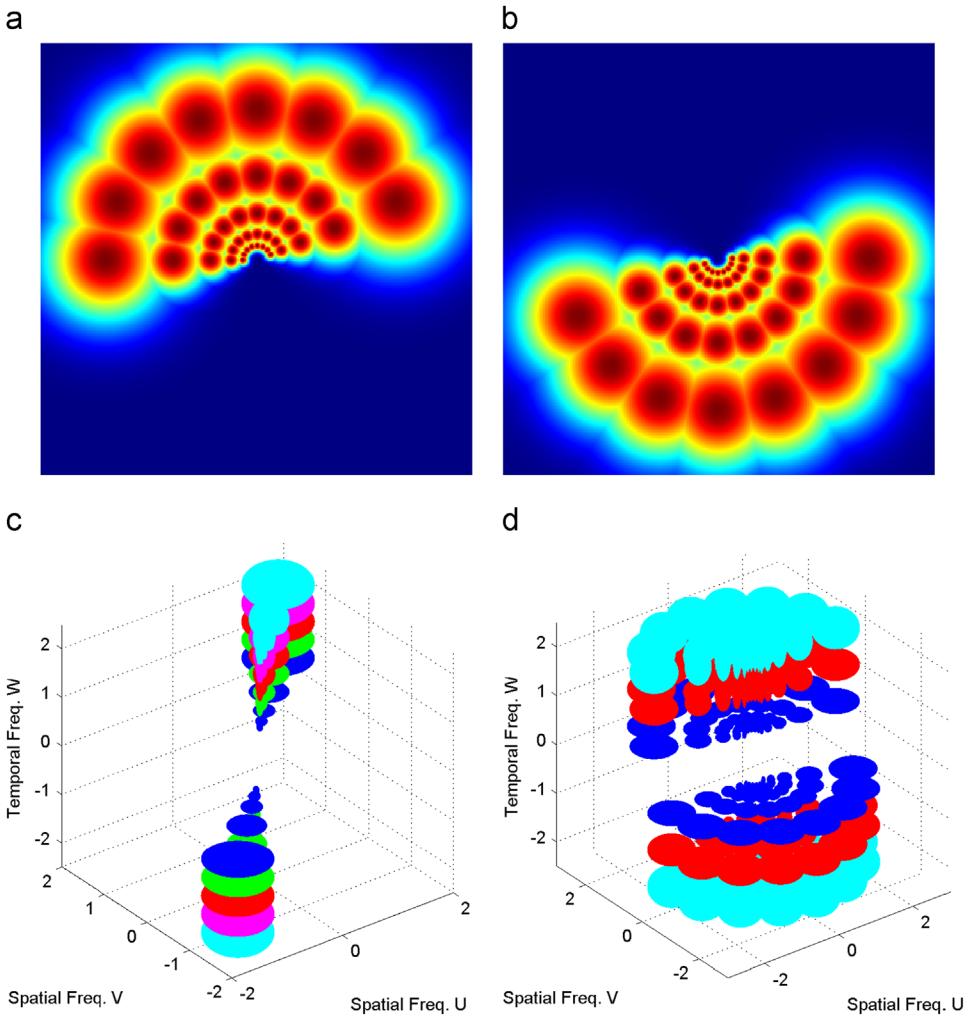

Fig. 3. Isosurfaces of the 3D spatio-temporal filterbank and a top view of a filterbank slice designed at temporal frequency ω_{t_0}. Isosurfaces correspond at 70%-peak bandwidth magnitude while different colors are used for different temporal frequencies. We can see that the symmetric lobe of each filter appeared at the plane defined by the temporal frequency $-\omega_{t_0}$ in contrast with the 2D case. We also note that the bandwidth of each filter changes depending on the spatial scale and temporal frequency: (a) filterbank at ω_{t_0} (top view); (b) filterbank at $-\omega_{t_0}$ (top view); (c) spatio-temporal filterbank at 5 different spatial scales, 1 of 8 orientation and 5 temporal frequencies; and (d) spatio-temporal filterbank at 5 different spatial scales, 8 spatial orientations and 3 of 5 temporal frequencies.

Thus, the frequency response of an even (cos) Gabor filter consists of two Gaussian ellipsoids symmetrically placed at frequencies $(\omega_{x_0}, \omega_{y_0}, \omega_{t_0})$ and $(-\omega_{x_0}, -\omega_{y_0}, -\omega_{t_0})$. Fig. 3 shows isosurfaces of the 3D spatio-temporal filterbank as well as a top view of a filterbank slice designed at some temporal frequency ω_{t_0}. Note that the symmetric lobes of each filter appear at the plane defined by the temporal frequency $-\omega_{t_0}$ in contrast with the 2D case. So, if we want to cover the spatial frequency plane at each temporal frequency we must include in our filterbank both positive and negative temporal frequencies. Further, the bandwidth of each filter varies with the spatial scale and temporal frequency.

The 3D filtering is a time consuming process due to the complexity of all required 3D convolutions. However, Gabor filters are separable [25], which means that we can filter each dimension separately using an impulse response having the form (6). In this way, we apply only 1D convolutions instead of 3D, which increases the efficiency of the computations. Then the 3D output can be easily composed from 1D filtering outputs by using simple trigonometric properties in two steps (first 2D and then 3D). First, we compose the 2D spatial output from the impulse responses $g_c(x)$, $g_s(x)$, $g_c(y)$, $g_s(y)$ for both the even- and odd-phase filter (we show the equations only for the luminance modality; the procedure is the same for color stream)

$$y_c^{2D}(x, y, t) = (I_1(x, y, t) * g_c(x)) * g_c(y) - (I_1(x, y, t) * g_s(x)) * g_s(y) \quad (9)$$

$$y_s^{2D}(x, y, t) = (I_1(x, y, t) * g_s(x)) * g_c(y) + (I_1(x, y, t) * g_c(x)) * g_s(y) \quad (10)$$

Then the final 3D output corresponding to spatio-temporal filtering can be obtained by convolving the above 2D output with the 1D temporal impulse responses

$$y_c^{3D}(x, y, t) = y_c^{2D}(x, y, t) * g_c(t) - y_s^{2D}(x, y, t) * g_s(t) \quad (11)$$

$$y_s^{3D}(x, y, t) = y_c^{2D}(x, y, t) * g_s(t) + y_s^{2D}(x, y, t) * g_c(t) \quad (12)$$

For an image of size $n \times n \times n$ and a convolution kernel of $m \times m \times m$ the complexity is reduced from $O(n^3 \cdot m^3)$ that is required for 3D convolutions to $O(3n^3 \cdot m)$ that is required for three separable 1D convolutions. Moreover, our 3D filterbank is highly parallelized as the basic operation is a simple 1D convolution and each 3D filter can be computed independently. At this time, the total computational time is about 6–8 s per video frame using Matlab code without any optimization for speed up in a 4-core personal computer.

For the spatio-temporal filterbank we used $K=400$ Gabor filters (isotropic in the spatial components) which are arranged in five spatial scales, eight spatial orientations and ten temporal frequencies. The spatial scales and orientations are selected to cover a squared 2D frequency plane in a similar way to the design by Havlicek et al. [90]. Then both center frequencies and Gaussian bandwidths are divided by the spatial sampling frequencies in order to get discrete filters with normalized frequency parameters that can be directly applied at every image size. We note that this process can lead to anisotropic spatial Gabor filters for non-square images, although the original design includes isotropic filters.

We use 10 temporal Gabor filters, five at positive and five at negative center frequencies due to the 3D spectrum symmetries. These are linearly spaced to span the normalized frequency axis and each filter's half-peak octave bandwidth is 0.75 octaves. Gabor filters in the temporal domain can model time varying patterns in the video, in a similar way that 2D Gabor describe texture patterns in video, while the use of different temporal frequencies can detect motions that have different directions. Fig. 3 shows spatio-temporal views of our design of this 3D filterbank. Note that including both positive and negative frequencies does not increase the filtering complexity because, due to Gabor filters' separability, no additional convolutions are needed but only changing the signs at (11) and (12). Finally, for the static (spatial only) filterbank we use the same spatial parameters with zero temporal frequency ($L=40$ filters), while for the lowpass filter we use both spatial and temporal zero frequencies. These three filterbank types can generate different features which play an important role in estimating spatio-temporal visual saliency. The spatio-temporal filterbank can detect motion activities, while the static one can find significant image regions which may attract human attention such as specific texture or strong edges. The low-pass filter can be related to what many models refer to as "intensity conspicuity" and describes video regions that have high values of luminance or color stream inside a spatio-temporal window (defined by the 3D Gaussian bandwidth).

3.2.2. Postprocessing

After the filtering process, for each filter i we obtain a quadrature pair output $(y_s^{3D}(x, y, t), y_c^{3D}(x, y, t))$ which corresponds to the even- and odd-phase 3D filter outputs. For each filter we can compute the total Gabor energy $E(\cdot)$, which is invariant to the phase of the input, by taking the sum of the squared energy of these two outputs

$$E(y_s^{3D}, y_c^{3D}) = \left(y_s^{3D}(x, y, t)\right)^2 + \left(y_c^{3D}(x, y, t)\right)^2 \quad (13)$$

After applying the above energy operator to each filter we have $K=400$ *energy volumes* for the spatio-temporal part (STE_i), $L=40$ for the static part (SE_i) and one for the low-pass filter (LE_0); see Fig. 2. In order to form one volume for each of these three independent filtering parts we apply the first step of *Dominant Component Analysis*[1] both to spatio-temporal and static energy volumes. Specifically, for each voxel (x, y, t) we keep its maximum value between all existing energy volumes

$$STDE = \max_{1 \le i \le K} STE_i, \quad SDE = \max_{1 \le i \le L} SE_i \quad (14)$$

Instead of keeping only the dominant energy we can keep the N highest spatio-temporal energies for each voxel and afterwards compute the average or the min value of them. This makes our analysis more robust to noise but requires N times more memory and storage space. In our experiments N assumes values in the range $[1, 6]$. For the low-pass energy we apply a simple center-surround difference in order to enhance regions which have significantly different values from their background. At each voxel of the video segment we subtract from its low-pass energy value $(LE_0(x, y, t))$ the average value of the entire low-pass energy volume

$$LE(x, y, t) = \left|LE_0(x, y, t) - \overline{LE}_0(x, y, t)\right| \quad (15)$$

Finally, we have three raw energy volumes for each luminance and color stream: spatio-temporal dominant energy *STDE* (see Figs. 5b and e and 6b and e), static dominant energy *SDE* (see Figs. 5c and f and 6c and f) and lowpass energy *LE* (see Figs. 5d and g and 6d and g). These raw energy volumes can be used as feature pools for composing a single saliency map under different fusion schemes.

Alternatively, these energy volumes can become further smoothed by applying a *temporal moving average* (TMA). Thus, each frame energy is computed as the mean inside a temporal window which includes T successive frames whose total duration is 1 s. In this way, we integrate visual events which take place close in time, in a similar way that humans are believed to do. A spatial Gaussian smoothing can also be applied, in order to find more compact and dense energy regions.

[1] Dominant component analysis on the outputs of Gabor filterbanks has been used for 2D texture analysis and segmentation in [91,90,92] and for spatio-temporal action classification in [93,94]. It may include additional steps of demodulation.

4. Spatio-temporal visual saliency

With the above described visual frontend, we obtain six energy volumes which quantitatively describe different aspects of visual saliency and thus provide a spatio-temporal feature set. In order to obtain a single spatio-temporal saliency map we combine these volumes using different fusion schemes. These can be categorized based on the applied fusion method or the specific feature subset used.

For fusion we have experimented with three widely used functions: min, max, mean, which correspond to different approaches in feature integration. Using the max we search for video regions that are salient in at least one energy map, the min keeps as salient those voxels which have high energy value in all energy volumes, whereas the mean lies somewhere in the middle. In the first step, we apply the selected fusion function to the dynamic volumes (luminance *STDE* and color *STDE*) and static volumes (luminance *SDE*, color *SDE*, luminance *LE* and color *LE*) independently. Then, we normalize both resulting volumes in the range $[0, 1]$. Finally, we fuse the normalized static and dynamic volumes by employing the same function used in the first step.

We have selected five feature subsets which are defined by which energy volumes we keep: (1) only luminance *STDE*, (2) luminance *STDE* and color stream *STDE*, (3) luminance *STDE* and color *LE*, (4) luminance and color stream *STDE*, luminance and color stream *SDE*, (5) all six energy volumes. The motivation behind these choices is that the luminance *STDE* describes the motion better, which is important for spatio-temporal saliency, while color information can be integrated both dynamically (*STDE*) or statically (*LE*). The use of all six energies means that both luminance and color stream are integrated into the final saliency map in a spatio-temporal (Lum.*STDE*, Col.*STDE*) as well as a spatially or temporally static (Lum.*SDE*, Col.*SDE*, Lum.Low, Col.Low) way.

For the qualitative evaluation of our spatio-temporal frontend we have created several simple stimuli where a time-varying event takes place among other spatial salient objects. In Fig. 4 we see two examples of these spatio-temporal stimuli. We also provide the saliency maps created by three state-of-the-art spatial saliency models (AWS [52], GBVS [48], Hou et al. [66]) as well as our spatio-temporal dominant energy (*STDE*). The general static background consists of some green and blue circles among many red circles. In stimulus A we have included a

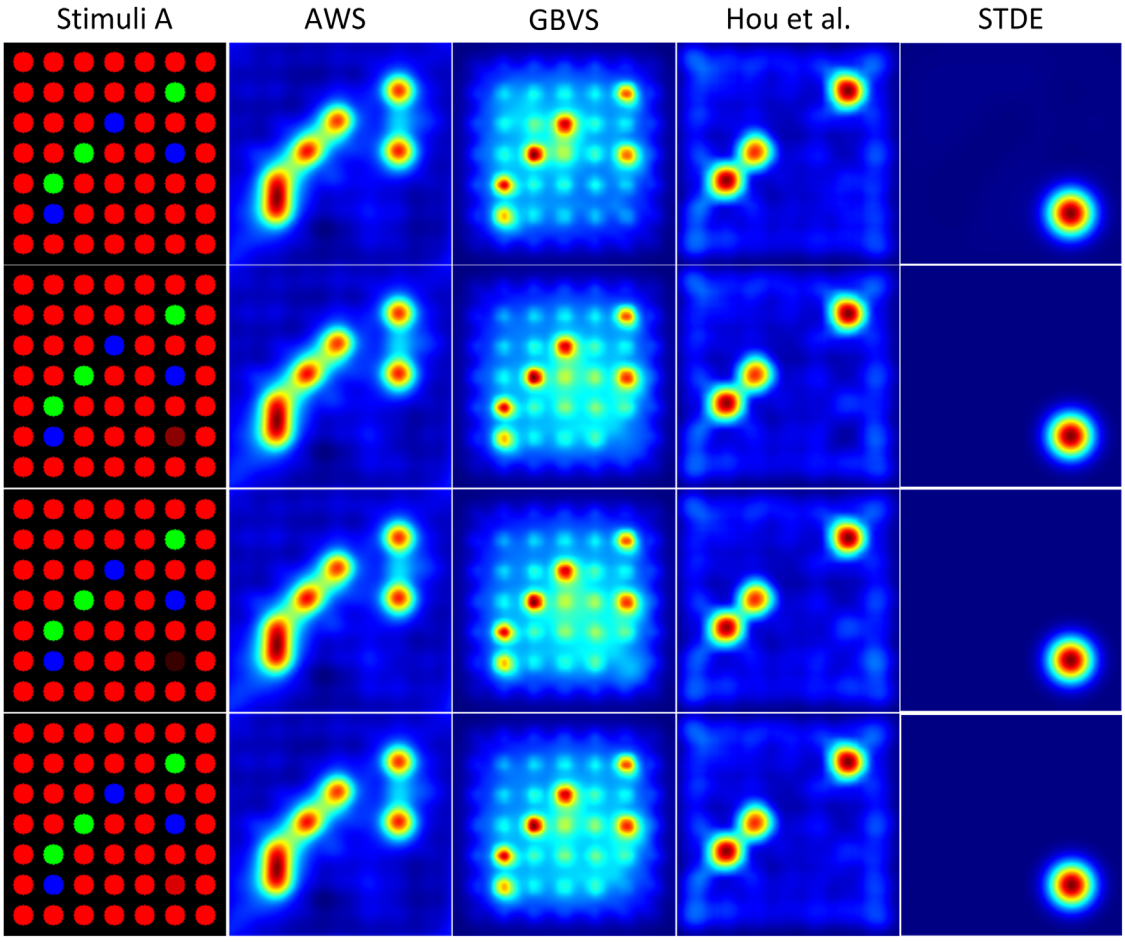

Fig. 4. Two synthetic spatio-temporal stimuli with the saliency maps created by three state-of-the-art static saliency models. We also see our *STDE* that detects the spatio-temporal events: (a) stimuli A and (b) stimuli B.

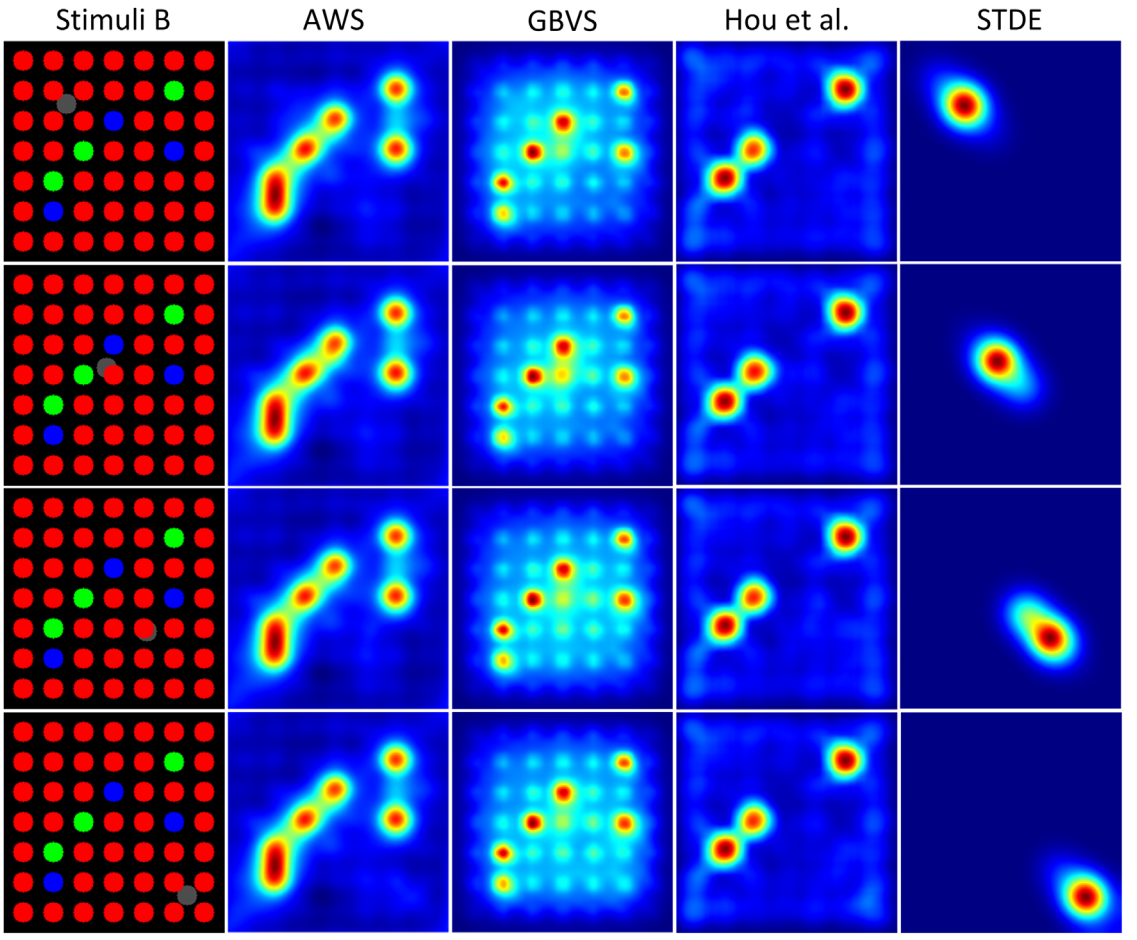

Fig. 4. (*continued*)

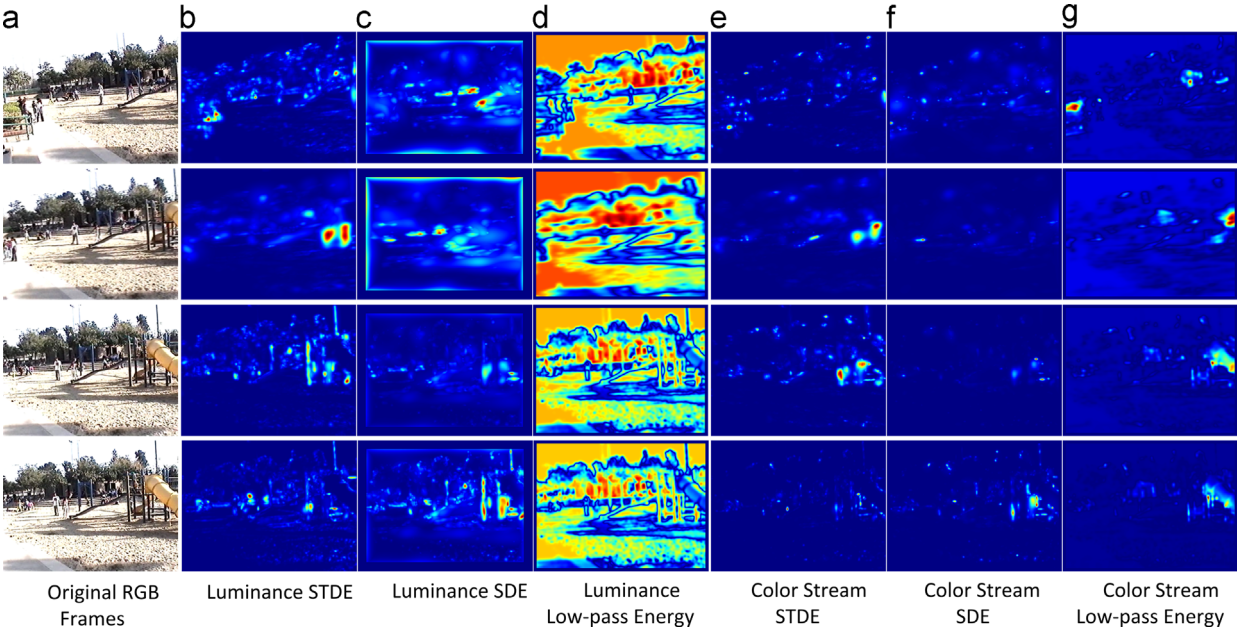

Fig. 5. Example frames of the six energy volumes computed using our frontend on the video *beverly01* from CRCNS-ORIG database. The beams of the *slide* on the right is detected by both luminance and color *STDE*, while the yellow "slide" is detected by only the Low-pass color energy.

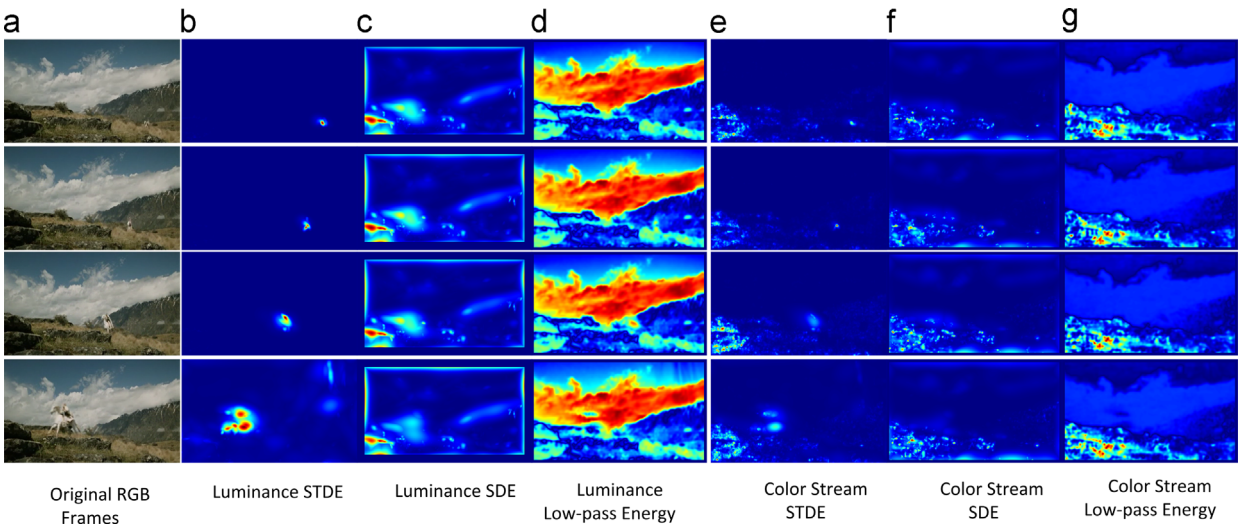

Fig. 6. Example frames of the six energy volumes computed using our frontend on the *Lord of the Rings* (Clip 1) from our Eye-Tracking Movie Database (ETMD). The galloping horse is perfectly detected by the luminance *STDE*.

blinking red circle whereas stimulus B contains a moving gray circle with low contrast. We see that all spatial saliency models detect as salient objects only the green and blue circles since they do not contain any temporal information. On the other hand our approach, which includes *STDE*, can detect both the spatial and the two time-varying events due to the multi-scale temporal filtering.

4.1. Evaluation on CRCNS-ORIG database

In order to quantitatively evaluate our proposed saliency estimation model we employ the widely used database CRCNS-ORIG by Itti et al. [9], which contains 50 short-length color video (about 25 minutes total playtime) with human eye-tracking annotation. Fig. 5 shows example frames of all six energies computed using our frontend from the video *beverly01* of the CRCNS-ORIG database. We note that the beams of the “slide”, which is introduced on the screen right after camera movement, are detected by both luminance and color *STDE* as these could be considered as spatio-temporal texture patterns. On the other hand the yellow “slide” is detected only by the low-pass color energy because, despite the fact that it moves, it retains a flat structure in the spatial domain. So this region is filtered by the spatial part of the 3D Gabor filter. The luminance *SDE*, as we can also see, models static texture patterns or edges while the luminance low-pass energy detects regions that have higher absolute luminance in relation to the frame's average.

We have tried to keep the same evaluation framework as in [82]. We compared our results according to the three evaluation scores, as they are described in [82]: Correlation Coefficient, Normalized Scanpath Saliency, Area Under Curve. Despite the spatio-temporal character of our method these three measures are computed at each frame separately.

Correlation Coefficient (CC) expresses the relationship between the model's saliency map and the saliency map created by centering a 2D gaussian at each viewer's eye fixation. *Normalized Scanpath Saliency* (NSS) is computed on the model's saliency map, after zero mean normalization and unit standardization, and shows how many times over the whole map's average is the model's saliency value at each human fixation. For NSS computation we subtract from the saliency map its average value and then divide with its standard deviation. Then the values of this normalized saliency map at each viewer fixation position consist the NSS values. As final NSS value we take the mean over all viewers fixations, while a negative NSS shows that the model cannot predict saliency region better than random selection. *Area Under Curve* (AUC) is defined by the area under the receiver operating characteristic (ROC) curve [95]. For our evaluation we consider saliency as a binary classification problem, in which saliency regions are included in the positive class while non-salient pixels form the negative set. Model's saliency values are the single features. After thresholding these values we take an ROC curve and subsequently the AUC measure. Instead of selecting the negative points uniformly from a video frame we use the *shuffled AUC* [56], which can be more robust across center-bias issue. According to shuffled AUC, we select the negative points from the union of all viewers' fixations across all other frames except the frame for which we compute the AUC. For more details about the above evaluation scores the reader is referred to [82,1].

The results for the different fusion functions and feature subsets using both raw and TMA energies are shown in Tables 1 and 2. We see that the raw energies perform in general better than the TMA. The mixture containing all energies gives the highest value for AUC score, while mixtures including luminance *STDE* and color low-pass yield large values w.r.t. CC and NSS. The fusion schemes using the mean have better performance and achieve higher values for the shuffled AUC. On the other hand, the min-based scheme for the luminance *STDE* and color low-pass feature mixture gives the best CC and NSS scores. In addition, the fusion schemes including only the dynamic volumes *STDE* can achieve a fairly good performance,

Table 1
Evaluation scores for the CRCNS-ORIG database using *LAB* color space.

Evaluation score		Correlation coefficient (CC) Fusion function			Normalized scanpath saliency (NSS) Fusion function			Shuffled area under curve (AUC) Fusion function		
Features		MEAN	MAX	MIN	MEAN	MAX	MIN	MEAN	MAX	MIN
Energy type	Feature subsets									
Raw	Lum.STDE	0.103	–	–	0.895	–	–	0.570	–	–
Raw	Lum.STDE/Col.STDE	0.097	0.092	0.091	0.840	0.801	0.789	**0.572**	0.571	0.561
Raw	Lum.STDE/Col.Low	0.101	0.084	**0.108**	0.885	0.731	**0.935**	0.569	0.557	0.565
Raw	STDE/SDE	0.077	0.062	0.073	0.664	0.536	0.627	**0.572**	0.561	0.554
Raw	All 6 energies	0.094	0.058	0.085	0.808	0.501	0.735	0.571	0.541	0.558
TMA	Lum.STDE	0.103	–	–	0.894	–	–	0.567	–	–
TMA	Lum.STDE/Col.STDE	0.096	0.093	0.088	0.828	0.799	0.762	**0.568**	0.567	0.556
TMA	Lum.STDE/Col.Low	0.105	0.088	**0.107**	0.905	0.762	**0.920**	0.566	0.556	0.560
TMA	STDE/SDE	0.078	0.065	0.068	0.670	0.561	0.585	**0.568**	0.559	0.549
TMA	All 6 energies	0.095	0.063	0.080	0.813	0.537	0.685	0.567	0.544	0.553

Table 2
Evaluation scores for the CRCNS-ORIG database using *PCA transformed* color space.

Evaluation score		Correlation coefficient (CC) Fusion function			Normalized scanpath saliency (NSS) Fusion function			Shuffled area under curve (AUC) Fusion function		
Features		MEAN	MAX	MIN	MEAN	MAX	MIN	MEAN	MAX	MIN
Energy type	Feature subsets									
Raw	Lum.STDE	0.112	–	–	0.970	–	–	0.578	–	–
Raw	Lum.STDE/Col.STDE	0.119	0.112	0.118	1.037	0.971	1.026	0.594	0.591	0.585
Raw	Lum.STDE/Col.Low	0.120	0.102	**0.126**	1.044	0.888	**1.096**	0.588	0.582	0.572
Raw	STDE/SDE	0.098	0.083	0.096	0.850	0.717	0.829	0.598	0.583	0.576
Raw	All 6 energies	0.117	0.081	0.114	1.006	0.699	0.989	**0.601**	0.575	0.580
TMA	Lum.STDE	0.111	–	–	0.955	–	–	0.574	–	–
TMA	Lum.STDE/Col.STDE	0.118	0.111	0.117	1.024	0.958	1.011	0.590	0.586	0.582
TMA	Lum.STDE/Col.Low	0.122	0.105	**0.126**	1.059	0.909	**1.089**	0.586	0.581	0.567
TMA	STDE/SDE	0.100	0.087	0.092	0.861	0.748	0.788	0.594	0.580	0.570
TMA	All 6 energies	0.119	0.086	0.109	1.020	0.737	0.941	**0.598**	0.578	0.574

partly because in video viewing temporal information has a significantly greater influence to the human attention than the static components. Moreover, we see that the use of the PCA transformed color space gave better results than the LAB color space. Despite the fact that the LAB color space is perceptually inspired, the PCA gives uncorrelated color streams adapted to each image content. In addition, there are physiological evidences that many processes in the human brain are closely related with the decorrelation of the input stimulus.

In Table 3 we see results after employing different functions in our dominant energy analysis. Specifically, we kept the $N=4$, 6 more dominant energy volumes and then we applied three different functions: mean, max, min. We have also used our best fusion scheme: mean fusion function, all energies as feature subset and raw energies. We see improvement to all evaluation metrics for both the mean and the min function. The results for the max function are by default the same regardless the number of dominant energies. So, taking the min (or mean) among N more dominant energies may be more robust to noise and computational errors than the classic dominant analysis but it requires more computational resources as well. In addition, the use of $N=6$ volumes seems to yield an additional increment to the results when we use the min function.

In order to compare our proposed method with other methods we have evaluated 15 state-of-the-art methods with publicly available code. In our comparisons we have included three spatio-temporal models that are related with the three basic approaches for visual saliency: (1) cognitive inspired [45,48], (2) statistical framework [64] and (3) frequency domain analysis [71,68]. Their results are presented in Table 4 together with our method's version that includes: dominant energy analysis using max function, PCA approach for the color stream, mean fusion function, all six energies as feature subset, raw energies and a gaussian spatial smoothing. Table 4 also presents the scores achieved by a Gaussian blob centered at the center of the image.

According to our evaluation results our visual frontend outperforms the other saliency estimation methods. Some of them are cognitively inspired [44,2] or use recent ideas about visual saliency, such as information theory measures

Table 3
Evaluation Scores for the CRCNS-ORIG database. In the comparison the following are included: *three different functions (mean, max, min) applied at N = 4, 6 more dominant energy volumes*, PCA color space, mean fusion function, all six energies as feature subset and raw energies.

Evaluation score	Correlation coefficient (CC) Applied function			Normalized scanpath saliency (NSS) Applied function			Shuffled area under curve (AUC) Applied function		
Number of dominant energies (N)	MEAN	MAX	MIN	MEAN	MAX	MIN	MEAN	MAX	MIN
$N=4$	0.118	0.117	0.121	1.022	1.006	1.040	0.604	0.601	0.605
$N=6$	0.119	0.117	**0.122**	1.029	1.006	**1.053**	0.604	0.601	**0.607**

Table 4
Comparison with state-of-the-art methods in the CRCNS-ORIG database. In our method are included: dominant energy analysis using min function, PCA color space, mean fusion, all six energies as feature subset and raw energies.

Method's name/citation	Spatial or spatio-temporal	Learning	Correlation coefficient (CC)	Normalized scanpath saliency (NSS)	Shuffled area under curve (AUC)
Our method	Spatio-temporal	NO	0.122	1.053	0.607
Our method + AWS	Spatio-temporal	NO	0.128	1.110	**0.621**
AIM: Bruce and Tsotos [58]	Spatial	YES	0.106	0.900	0.598
AWS: Diaz et al. [52]	Spatial	NO	0.108	0.936	0.608
GBVS: Harel et al. [48]	Spatial	NO	0.169	1.454	0.574
GBVS: Harel et al. [48]	Spetio-temporal	NO	**0.173**	**1.504**	0.590
Hou and Zhang [66]	Spatial	NO	0.108	0.930	0.603
Itti et al. 1 [44,48]	Spatial	NO	0.122	1.039	0.565
Itti et al. 1 [45,48]	Spetio-temporal	NO	0.131	1.123	0.582
Itti et al. 2 [44,2]	Spatial	NO	0.056	0.525	0.527
Itti et al. 2 [45,2]	Spetio-temporal	NO	0.084	0.795	0.547
Judd et al. [75]	Spatial	YES	0.165	1.401	0.565
PQFT: Guo et al. [71,68]	Spetio-temporal	NO	0.118	1.059	0.590
SDSR: Seo and Milanfar [64]	Spetio-temporal	NO	0.104	0.904	0.584
SUN: Zhang et al. [56]	Spatial	YES	0.075	0.633	0.573
Torralba [53]	Spatial	NO	0.091	0.779	0.585
Gaussian Blob [82]	Spatial	NO	0.152	1.268	0.500

[53,58,56], ICA analysis with precomputed basis using datasets of natural images [58,56] and saliency estimation in the frequency domain [66], but they do not process the videos in the temporal direction as they are originally designed for static images. Our proposed method also outperforms the three state-of-the-art spatio-temporal models that include a temporal information channel but estimate motion by computing differences between 2D orientation maps of successive frames [45,48] or using a small number of (e.g. 2–3) frames, instead of applying spatio-temporal filtering at different scales and orientations as in our proposed frontend. Regarding the shuffled AUC score, which is robust across center-bias issue, our method scores as well as the AWS method. On the other hand, our method outperforms the AWS w.r.t. CC and NSS scores. The AWS method uses spatial 2D log-Gabor filters and decorrelates the multiscale filter responses using PCA analysis [52] in order to obtain the final saliency map. In our method, we use simpler ideas like dominant energy analysis and separable Gabor filters and process the video volume simultaneously in three dimensions (spatial and temporal). The combination of our method with the static AWS method gives a significant improvement in the AUC score.

We have to note that many of the methods we have evaluated are designed and used for still images. Therefore applying them at each video frame in video sequences becomes a time consuming process in most cases (e.g. [75]). Our proposed method is designed for videos and is purely bottom up since it is based on perceptually inspired spatio-temporal features and uses simple and fast fusion techniques instead of machine learning or other advanced and complex methods for feature integration. Finally, we note that the employed evaluation measures, which were widely used in the literature, are designed for static images; so the frame by frame evaluation of the visual saliency in a video stimuli is not always the best choice.

4.2. Eye-Tracking Movie Database (ETMD)

In our general effort to deal with the movie summarization problem it would be useful if we could evaluate our visual saliency model on a database that contains longer and more complex video clips. For this reason, we have developed a database comprising video clips from Hollywood movies which we have enriched with eye-tracking human annotation: the Eye-Tracking Movie Database (ETMD). Specifically, we cut two short video clips (about 3–3.5 min) from each one of six Oscar-winning movies of

various genres: Chicago, Crash, Departed, Finding Nemo, Gladiator, Lord of the Rings – The return of the King. We tried to include scenes with high motion and action as well as dialogues. These clips were annotated with eye-tracking data by 10 different people (annotation data from at least eight people were collected for each clip). We asked them to see in full screen these clips both in grayscale and in color, while an eye-tracking system recorded their eyes fixations on the screen. Specifically, we have used the commercial Eye Tracking System TM3 provided by Eye-TechDS. This device uses a camera with infrared light and provides a real time continuous gaze estimation, defined as fixation points on the screen. The tracker's rate has been selected to be synchronized with the video frame rate in order to have one fixation point pair per frame. For visual saliency problems a weighted average between two eye fixations is provided, which is defined either by the mean, if both eyes are found by the eye-tracker, or only by the detected eye's fixation. If neither eye is detected or the fixations lie out of screen boundaries, fixation gets a zero value. Fig. 7 shows examples of the fixation points at frame #500 for each of the 12 movie clips. We see that in most cases the fixation points of all viewers lie in general close to each other. By analyzing the eye-tracking data we provide in Table 5 useful statistics regarding the database, such as the number of frames, total duration and valid fixation points per frame, and find correlations among the different viewers and between the color and grayscale version of each movie clip. We see that the fixations are generally correlated both between the different users and the version (color or grayscale) of each movie clip. However, in some movies, such as CHI, the fixations data are highly correlated while other clips (FNE Clip 2, LOR Clip 2) have lower correlation values.

We have applied and evaluated our computational model on this novel database in a similar way as we did for the CRCNS-ORIG database. Fig. 6 shows example frames of all six model's energies computed on the video *Lord of the Rings* (Clip 1) from our new Eye-Tracking Movie Database (ETMD). We note that the white galloping horse is perfectly detected by only the luminance *STDE*, since its color information is negligible. The Luminance/color *SDE* and Low-pass energies model static objects or regions in the video sequence, like the rock in the bottom-left and the clouds in the air.

The results for the different fusion schemes using the PCA transformed color space and the both Raw and TMA energies are shown in Table 6. Here, we keep the minimum of the $N=6$ more dominant spatio-temporal energy volumes. There is also one additional subset of features: luminance *STDE* evaluated on grayscale annotated clips.

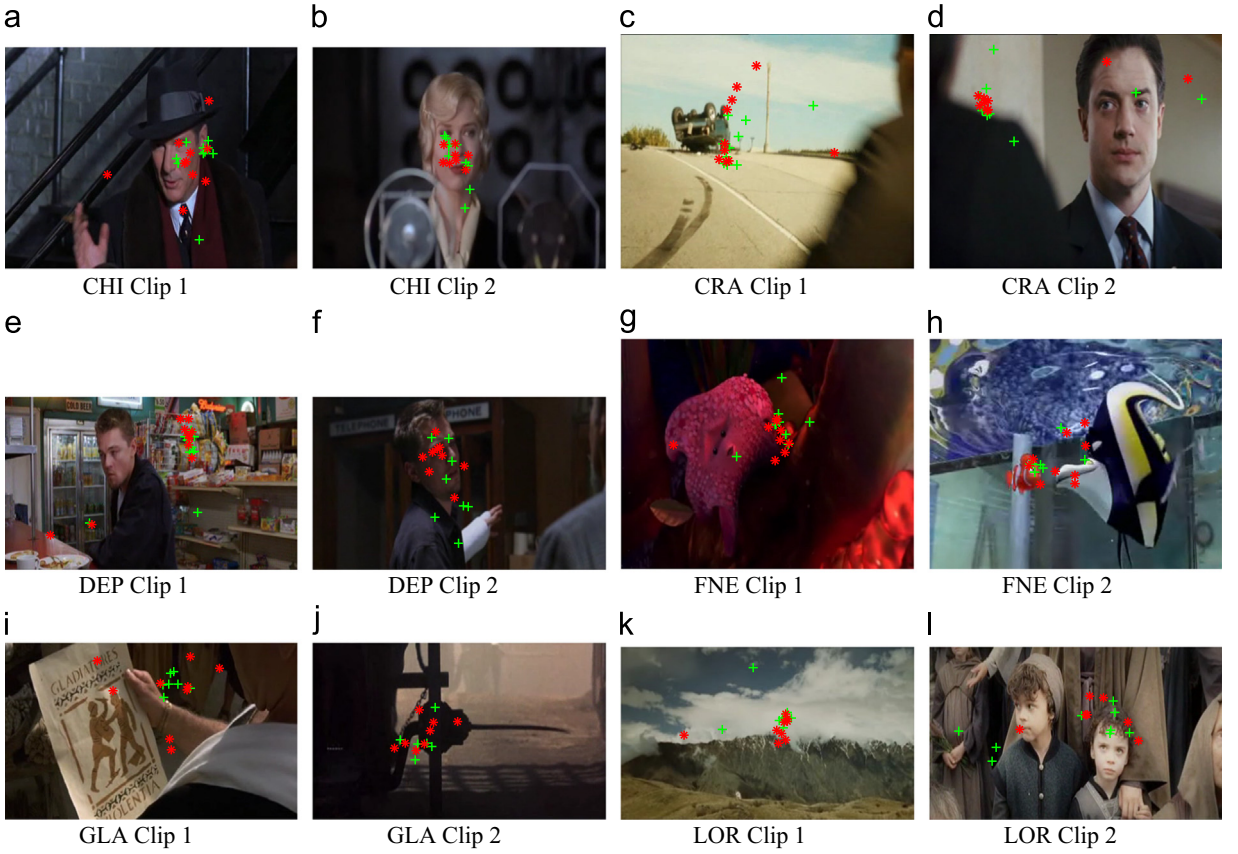

Fig. 7. Examples of the fixation points at frame no. 500 for each of the 12 movie clips. With green + are the fixations points over the color version of each clip, while with red ∗ are the points for the grayscale version. Best viewed in color: (a) CHI Clip 1; CHI Clip 2; (c) CRA Clip 1; (d) CRA Clip 2; (e) DEP Clip 1; (f) DEP Clip 2; (g) FNE Clip 1; (h) FNE Clip 2; (i) GLA Clip 1; (j) GLA Clip 2; (k) LOR Clip 1; and (l) LOR Clip 2.

Table 5
Statistics for the Eye Tracking Movie Database (ETMD).

Video clip name	Number of frames	Duration (min)	Number of viewers	Valid fixations number per frame	Average correlation between viewers	Average correlation between color and grayscale version
CHI Clip 1	5075	03:22	10	9.50	0.506	0.495
CHI Clip 2	5241	03:29	9	8.63	0.430	0.484
CRA Clip 1	5221	03:28	10	9.47	0.335	0.310
CRA Clip 2	5079	03:23	9	8.47	0.406	0.467
DEP Clip 1	4828	03:13	10	9.45	0.520	0.548
DEP Clip 2	5495	03:39	9	8.25	0.473	0.534
FNE Clip 1	5069	03:22	9	8.45	0.372	0.371
FNE Clip 2	5083	03:23	8	7.50	0.292	0.294
GLA Clip 1	5290	03:31	9	8.18	0.423	0.407
GLA Clip 2	4995	03:19	8	7.61	0.354	0.443
LOR Clip 1	5116	03:24	9	8.38	0.452	0.431
LOR Clip 2	5152	03:26	8	7.56	0.294	0.283

Table 6
Evaluation Scores for the Eye-Tracking Movie Database (ETMD) using *PCA transformed* color space. The evaluation of the luminance STDE was based on eye-tracking annotation on both a grayscale and color version of each video.

Evaluation score		Correlation coefficient (CC) Fusion function			Normalized scanpath saliency (NSS) Fusion function			Shuffled area under curve (AUC) Fusion function		
Features		MEAN	MAX	MIN	MEAN	MAX	MIN	MEAN	MAX	MIN
Energy type	Feature subsets									
Raw	Lum.STDE (Grayscale Annot.)	0.113	–	–	0.786	–	–	0.603	–	–
Raw	Lum.STDE (Color Annot.)	0.114	–	–	0.794	–	–	0.601	–	–
Raw	Lum.STDE/Col.STDE	0.121	0.111	0.127	0.837	0.765	0.886	0.618	0.610	0.615
Raw	Lum.STDE/Col.Low	0.145	0.127	0.149	1.016	0.892	1.032	0.630	0.613	0.616
Raw	STDE/SDE	0.110	0.093	0.115	0.761	0.640	0.797	0.620	0.604	0.608
Raw	All 6 energies	0.128	0.097	0.135	0.886	0.672	0.931	0.629	0.606	0.616
TMA	Lum.STDE (Grayscale Annot.)	0.120	–	–	0.827	–	–	0.611	–	–
TMA	Lum.STDE (Color Annot.)	0.120	–	–	0.835	–	–	0.609	–	–
TMA	Lum.STDE/Col.STDE	0.130	0.119	0.136	0.897	0.820	0.944	0.629	0.620	0.627
TMA	Lum.STDE/Col.Low	0.156	0.135	**0.162**	1.083	0.944	**1.124**	**0.643**	0.625	0.629
TMA	STDE/SDE	0.119	0.101	0.120	0.816	0.695	0.828	0.630	0.613	0.616
TMA	All 6 energies	0.138	0.106	0.142	0951	0.732	0.977	0.641	0.616	0.625

Our goal for this addition was to also evaluate our method without the influence of the color stream.

We see that according to all 3 metrics the TMA energies perform significantly better than the Raw in this new dataset. They give smoother energies and so the salient events in the movies can be determined more accurately. We can see that the *STDE* and color low-pass feature subset has in general the best performance and the mean fusion performs quite better than the other schemes. The above results, regarding the TMA energies and the best feature and fusion scheme, are different than those we have obtained for the CRCNS-ORIG, in which the all energies mixture with Raw energies gave the best results. This could be explained as we consider that an object with high motion and color intensity may attract human attention more than the static parts of the scene. Moreover, regarding the grayscale versus color annotation, we saw that color based energy volumes can improve the performance over using only the luminance *STDE* and confirms the need for incorporating color information in a saliency model. Thus, it seems that the luminance only *STDE* is adequate for the grayscale version of annotation, whereas the availability of color modeling clearly provides helpful additional information in the case of color annotation. This can be explained by the fact that despite the apparent correlation between the fixations points over the two versions (grayscale or color) of each clip, in many cases the viewers focus on different places in the video when the color information is on. We also note that the ETMD is a quite challenging database due to the existence of many shot and scene changes in the Hollywood movies. Finally, movies are highly face-biased, which means that during a movie the viewer mostly focuses on actors faces.

In order to compare our method's performance with the other state-of-the-art models we have evaluated the same visual saliency methods as in the CRCNS-ORIG. Their results are presented in Table 7 together with our best method's version that includes: dominant energy analysis using the minimum of the $N=6$ more dominant spatio-temporal energies, the PCA based color space, mean fusion function, luminance *STDE* and color low-pass energy as feature subset and TMA energies. We see that in this

Table 7
Comparison with state-of-the-art methods in the ETMD database. In our method are included: dominant energy analysis using min function, PCA color space, mean fusion, luminance *STDE* and color low-pass as feature subset and TMA energies.

Method's name/citation	Spatial or spatio-temporal	Learning	Correlation coefficient (CC)	Normalized scanpath saliency (NSS)	Shuffled area under curve (AUC)
Our method	Spatio-temporal	NO	0.156	1.083	**0.643**
AIM: Bruce and Tsotos [58]	Spatial	YES	0.138	0.919	0.610
AWS: Diaz et al. [52]	Spatial	NO	0.113	0.788	0.631
GBVS: Harel et al. [48]	Spatial	NO	0.213	1.435	0.598
GBVS: Harel et al. [48]	Spetio-temporal	NO	0.202	1.371	0.607
Hou et al. [66]	Spatial	NO	0.091	0.635	0.599
Itti et al. 1 [44,48]	Spatial	NO	0.166	1.093	0.549
Itti et al. 1 [45,48]	Spetio-temporal	NO	0.171	1.140	0.573
Itti et al. 2 [44,2]	Spatial	NO	0.058	0.429	0.542
Itti et al. 2 [45,2]	Spetio-temporal	NO	0.070	0.526	0.556
Judd et al. [75]	Spatial	YES	0.222	1.481	0.602
PQFT: Guo et al. [71,68]	Spetio-temporal	NO	0.095	0.688	0.558
SDSR: Seo and Milanfar [64]	Spetio-temporal	NO	0.084	0.587	0.574
SUN: Zhang et al. [56]	Spatial	YES	0.095	0.656	0.599
Torralba [53]	Spatial	NO	0.107	0.736	0.612
Gaussian Blob [82]	Spatial	NO	**0.239**	**1.569**	0.500

movie dataset our perceptually inspired spatio-temporal method yields a higher performance than any other evaluated state-of-the-art saliency model.

5. Conclusion

In this work we have dealt with the problem of spatio-temporal visual saliency with applications in eye-tracking annotated videos. We have proposed a new spatio-temporal computational visual frontend for estimating visual saliency. Our approach performed better than many other methods over the CRCNS-ORIG database, according to certain numerical criteria. It yields a quite high performance also for our new eye-tracking annotated movie database, which is an additional contribution of this paper. Our perceptually inspired spatio-temporal frontend employs simple fusion schemes on simply computed energy features, and its overall approach is low-level and bottom-up without any training process. As future work, we focus on the reduction of the frontend's complexity and integration of our bottom-up frontend with the movies' high level semantic information. Moreover, in our ongoing work we envision applications of this frontend to the movie summarization problem.

Acknowledgments

This research work was supported by the project “COGNIMUSE” which is implemented under the “ARISTEIA” Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund and Greek National Resources. It was also partially supported by the European Union under the project “MOBOT” with Grant FP7-600796.

The authors wish to thank all the members of the NTUA CVSP Lab who participated in the eye-tracking annotation of the movie database. Special thanks to Nassos Katsamanis for his advices during database collection and detailed comments on the paper.

Appendix A. Supplementary material

Supplementary data associated with this paper can be found in the online version at http://dx.doi.org/10.1016/j.image.2015.08.004. Additional information can be found at: http://cognimuse.cs.ntua.gr.

References

[1] A. Borji, L. Itti, State-of-the-art in visual attention modeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 185–207.
[2] D. Walther, C. Koch, Modeling attention to salient proto-objects, J. Neural Netw. 19 (9) (2006) 1395–1407.
[3] T. Kadir, M. Brady, Saliency, scale and image description, Int. J. Comput. Vis. 45 (2) (2001) 83–105.
[4] K. Rapantzikos, Y. Avrithis, S. Kollias, Dense saliency-based spatio-temporal feature points for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[5] Z. Liu, W. Zou, L. Li, L. Shen, O. le Meur, Co-saliency detection based on hierarchical segmentation, IEEE Signal Process. Lett. 21 (1) (2014) 88–92.
[6] Y. Ma, X. Hua, L. Lu, H. Zhang, A generic framework of user attention model and its application in video summarization, IEEE Trans. Multimed. 7 (5) (2005) 907–919.
[7] A. Money, H. Agius, Video summarization: a conceptual framework and survey of the state of the art, J. Vis. Commun. Image Represent. 19 (2) (2008) 121–143.
[8] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, Y. Avrithis, Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention, IEEE Trans. Multimed. 15 (7) (2013) 1553–1568, http://dx.doi.org/10.1109/TMM.2013.2267205.
[9] L. Itti, P. Baldi, Bayesian surprise attracts human attention, in: Proceedings of the NIPS, 2005.
[10] D.H. Hubel, T.N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, J. Physiol. 160 (1962) 106–154.

[11] D.H. Hubel, T.N. Wiesel, Receptive fields and functional architecture of monkey striate cortex, J. Physiol. 195 (1968) 215–243.
[12] J. Movshon, I. Thompson, D. Tolhurst, Spatial summation in the receptive fields of simple cells in the cat's striate cortex, J. Physiol. 283 (1978) 53–77.
[13] J. Movshon, I. Thompson, D. Tolhurst, Receptive field organization of complex cells in the cat's striate cortex, J. Physiol. 283 (1978) 79–99.
[14] R.L.D. Valois, E.W. Yund, N. Hepler, The orientation and direction selectivity of cells in macaque visual cortex, Vis. Res. 22 (1982) 531–544.
[15] F.W. Campbell, J.G. Robson, Application of Fourier analysis to the visibility of gratings, J. Physiol. (Lond.) 197 (1968) 551–566.
[16] L. Maffei, A. Fiorentini, The visual cortex as a spatial frequency analyzer, Vis. Res. 13 (1973) 1255–1267.
[17] H.R. Wilson, S.C. Giese, Threshold visibility of frequency gradient patterns, Vis. Res. 17 (1977) 1177–1190.
[18] J.G. Daugman, Two-dimensional spectral analysis of cortical receptive field profiles, Vis. Res. 20 (10) (1980) 847–856.
[19] D.A. Pollen, S.F. Ronner, Phase relationships between adjacent simple cells in the visual cortex, Science 212 (4501) (1981) 1409–1411.
[20] J. Daugman, Uncertainty relation for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical filters, J. Opt. Soc. Am. A 2 (7) (1985) 1160–1169.
[21] D. Gabor, Theory of communication, IEE J. (Lond.) 93 (1946) 429–457.
[22] S. Marcelja, Mathematical description of the responses of simple cortical cells, J. Opt. Soc. Am. 70 (11) (1980) 1297–1300.
[23] A.C. Bovik, M. Clark, W.S. Geisler, Multichannel texture analysis using localized spatial filters, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1) (1990) 55–73.
[24] E.H. Adelson, J.R. Bergen, Spatiotemporal energy models for the perception of motion, J. Opt. Soc. Am. A 2 (2) (1985) 284–299.
[25] D.J. Heeger, Model for the extraction of image flow, J. Opt. Soc. Am. 4 (8) (1987) 1455–1471.
[26] D.J. Heeger, Optical flow using spatio-temporal filters, Int. J. Comput. Vis. 1 (4) (1988) 279–302.
[27] H.R. Wilson, J.R. Bergen, A four mechanism model for spatial vision, Vis. Res. 19 (1979) 19–32.
[28] R.A. Young, Simulation of human retinal function with the gaussian derivative model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1986.
[29] R.A. Young, The Gaussian derivative model for spatial vision: I. Retinal mechanisms, Spat. Vis. 2 (4) (1987) 273–293.
[30] D. Marr, E. Hildreth, Theory of edge detection, Proc. R. Soc. Lond. B 207 (1980) 187–217.
[31] J.J. Koenderink, A. van Doorn, Representation of local geometry in the visual system, Biol. Cybern. 55 (1987) 367–375.
[32] R.A. Young, R.M. Lesperance, W.W. Meyer, The Gaussian derivative model for spatial–temporal vision: I. Cortical model, Spat. Vis. 14 (3,4) (2001) 261–319.
[33] M. Morrone, J. Ross, D. Burr, R. Owens, Mach bands depend on spatial phase, Nature 324 (1986) 250–253.
[34] M. Morrone, D. Burr, Feature detection in human vision: a phase-dependent energy model, Proc. R. Soc. Lond. B 235 (1988) 221–245.
[35] H. Knutsson, G.H. Granlund, Texture analysis using two-dimensional quadrature filters, in: Proceedings of the Workshop on Computer Architectures for Pattern Analysis and Image Database Management, 1983.
[36] J. Malik, P. Perona, Preatentive texture discrimination with early vision mechanisms, J. Opt. Soc. Am. A 7 (5) (1990) 923–932.
[37] D.R. Martin, C.C. Fowlkes, J. Malik, Learning to detect natural image boundaries using local brightness, color, and texture cues, IEEE Trans. Pattern Anal. Mach. Intell. 26 (5) (2004) 530–549.
[38] A. Treisman, G. Gelade, A feature integration theory of attention, Cognit. Psychol. 12 (1) (1980) 97–136.
[39] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Human Neurobiol. 4 (4) (1985) 219–227.
[40] R. Milanese, Detecting salient regions in an image: from biological evidence to computer implementation (Ph.D. thesis), University of Geneva, 1993.
[41] S. Baluja, D. Pomerleau, Using a saliency map for active spatial selective attention: implementation & initial results, in: Proceedings of the NIPS, 1994.
[42] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis, F. Nuflo, Modeling visual attention via selective tuning, Artif. Intell. 78 (1–2) (1995) 507–545.
[43] E. Niebur, C. Koch, Control of selective visual attention: Modeling the where pathway, in: Proceedings of the NIPS, 1995.
[44] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.
[45] L. Itti, N. Dhavale, F. Pighin, Realistic avatar eye and head animation using a neurobiological model of visual attention, in: Proceedings of the SPIE 48th Annual International Symposium on Optical Science and Technology, 2003.
[46] S. Frintrop, VOCUS: a visual attention system for object detection and goal-directed search, Lecture Notes in Computer Science, vol. 3899, Springer, Berlin 2006.
[47] O.L. Meur, P.L. Callet, D. Barba, D. Thoreau, A coherent computational approach to model bottom-up visual attention, IEEE Pattern Anal. Mach. Intell. 28 (5) (2006) 802–817.
[48] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: NIPS, 2006.
[49] O.L. Meur, P.L. Callet, D. Barba, Predicting visual fixations on video based on low-level visual features, Vis. Res. 47 (19) (2007) 2483–2498.
[50] S. Marat, T. Ho-Phuoc, L. Granjon, N. Guyader, D. Pellerin, A. Guérin-Dugué, Modelling spatio-temporal saliency to predict gaze direction for short videos, Int. J. Comput. Vis. 82 (3) (2009) 231–243.
[51] K. Rapantzikos, Y. Avrithis, S. Kollias, Spatiotemporal features for action recognition and salient event detection, Cognit. Comput., Special Issue on Saliency, Attention, Visual Search and Picture Scanning 3 (1) (2011) 167–184.
[52] A. Garcia-Diaz, X.R. Fernandez-Vidal, X.M. Pardo, R. Dosil, Saliency from hierarchical adaptation through decorrelation and variance normalization, Image Vis. Comput. 30 (1) (2012) 51–64.
[53] A. Torralba, Modeling global scene factors in attention, J. Opt. Soc. Am. A 20 (2003) 1407–1418.
[54] I. Gkioulekas, G. Evangelopoulos, P. Maragos, Spatial Bayesian surprise for image saliency and quality assessment, in: Proceedings of the International Conference on Image Processing, 2010.
[55] A. Oliva, A. Torralba, M.S. Castelhano, J.M. Henderson, Top-down control of visual attention in object detection, in: Proceedings of the International Conference on Image Processing, 2003.
[56] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, G.W. Cottrell, Sun: a Bayesian framework for saliency using natural statistics, J. Vis. 8 (7) . 32.1–20.
[57] L. Zhang, M.H. Tong, G.W. Sunday, Saliency using natural statistics for dynamic analysis of scenes, in: Proceedings of the Thirty-first Annual Cognitive Science Society Conference, 2009.
[58] N. Bruce, J. Tsotsos, Saliency based on information maximization, in: Proceedings of the NIPS, 2005.
[59] N.D.B. Bruce, J.K. Tsotsos, Spatiotemporal saliency: towards a hierarchical representation of visual saliency, in: International Workshop on Attention and Performance in Computer Vision, 2008.
[60] X. Hou, L. Zhang, Dynamic visual attention: searching for coding length increments, in: NIPS, 2009.
[61] D. Gao, N. Vasconcelos, Discriminant saliency for visual recognition from cluttered scenes, in: Proceedings of the NIPS, 2004.
[62] D. Gao, S. Han, N. Vasconcelos, Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (6) (2009) 989–1005.
[63] V. Mahadevan, N. Vasconcelos, Spatiotemporal saliency in dynamic scenes, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2010) 171–177.
[64] H.J. Seo, P. Milanfar, Static and space-time visual saliency detection by self-resemblance, J. Vis. 9 (12) (2009) 15.
[65] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, T. Dutoit, Rare2012: a multi-scale rarity-based saliency detection with its comparative statistical analysis, Signal Process.: Image Commun. 28 (6) (2013) 642–658.
[66] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[67] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[68] C. Guo, L. Zhang, A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression, IEEE Trans. Image Process. 19 (1) (2010) 185–198.
[69] X. Hou, J. Harel, C. Koch, Image signature: highlighting sparse salient regions, IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2012) 194–201.
[70] B. Schauerte, R. Stiefelhagen, Quaternion-based spectral saliency detection for eye fixation prediction, in: Proceedings of the European Conference on Computer Vision, 2012.
[71] C. Guo, Q. Ma, L. Zhang, Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[72] X. Cui, Q. Liu, D. Metaxas, Temporal spectral residual: fast motion saliency detection, in: Proceedings of the ACM International Conference on Multimedia, 2009.
[73] M. Mancas, N. Riche, J. Leroy, B. Gosselin, Abnormal motion selection in crowds using bottom-up saliency, in: Proceedings of the International Conference on Image Processing, 2011.
[74] T.V. Nguyen, M. Xu, G. Gao, M. Kankanhalli, Q. Tian, S. Yan, Static saliency vs. dynamic saliency: a comparative study, in: Proceedings of the ACM International Conference on Multimedia, 2013.
[75] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[76] W.-F. Lee, T.-H. Huang, S.-L. Yeh, H.H. Chen, Learning-based prediction of visual attention for video signals, IEEE Trans. Image Process. 20 (11) (2011) 3028–3038.
[77] D. Rudoy, D.B. Goldman, E. Shechtman, L. Zelnik-Manor, Learning video saliency from human gaze using candidate selection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[78] S.-H. Lee, J.-H. Kim, K.P. Choi, J.-Y. Sim, C.-S. Kim, Video saliency detection based on spatiotemporal feature learning, in: Proceedings of the International Conference on Image Processing, 2014.
[79] A. Borji, D.N. Sihite, L. Itti, Probabilistic learning of task-specific visual attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 470–477.
[80] J. Li, Y. Tian, T. Huang, W. Gao, Probabilistic multi-task learning for visual saliency estimation in video, Int. J. Comput. Vis. 90 (2) (2010) 150–165.
[81] A. Toet, Computational versus psychophysical image saliency: a comparative evaluation study, IEEE Trans. Pattern Anal. Mach. Intell. 33 (11) (2011) 2131–2146.
[82] A. Borji, D.N. Sihite, L. Itti, Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study, IEEE Trans. Image Process. 22 (1) (2013) 55–69.
[83] G. Wyszecki, W.S. Stiles, Color Science, 2nd ed. J. Wiley & Sons, New York, 1982.
[84] S. Gao, K. Yang, C. Li, Y. Li, A color constancy model with double-opponency mechanisms, in: Proceedings of the International Conference on Computer Vision, 2013.
[85] R.W. Rodieck, Quantitative analysis of cat retinal ganglion cell response to visual stimuli, Vis. Res. 5 (12) (1965) 583–601.
[86] K.G. Derpanis, M. Sizintsev, K. Cannons, R.P. Wildes, Efficient action spotting based on a spacetime oriented structure representation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[87] W.T. Freeman, E.H. Adelson, The design and use of steerable filters, IEEE Trans. Pattern Anal. Mach. Intell. 13 (6) (1991) 891–906.
[88] H. Takeda, S. Farsiu, P. Milanfar, Kernel regression for image processing and reconstruction, IEEE Trans. Image Process. 16 (2) (2007) 349–366.
[89] A. Belardinelli, F. Pirri, A. Carbone, Motion saliency maps from spatiotemporal filtering, in: Attention in Cognitive Systems, Springer, Berlin Heidelberg, 2009, pp. 112–123.
[90] J.P. Havlicek, D.S. Harding, A.C. Bovik, Multidimensional quasi-eigenfunction approximations and multicomponent am-fm models, IEEE Trans. Image Process. 9 (2) (2000) 227–242.
[91] A.C. Bovik, N. Gopal, T. Emmoth, A. Restrepo, Localized measurement of emergent image frequencies by Gabor wavelets, IEEE Trans. Inf. Theory 38 (1992) 691–712.
[92] I. Kokkinos, G. Evangelopoulos, P. Maragos, Texture analysis and segmentation using modulation features, generative models and weighted curve evolution, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 142–157.
[93] C. Georgakis, P. Maragos, G. Evangelopoulos, D. Dimitriadis, Dominant spatio-temporal modulations and energy tracking in videos: application to interest point detection for action recognition, in: Proceedings of the International Conference on Image Processing, 2012.
[94] K. Maninis, P. Koutras, P. Maragos, Advances on action recognition in videos using and interest point detector based on multiband spatio-temporal energies, in: Proceedings of the International Conference on Image Processing, 2014.
[95] D.M. Green, J.A. Swets, Signal Detection Theory and Psychophysics, Wiley, New York, 1966.