Multimed Tools Appl (2016) 75:7761–7778
DOI 10.1007/s11042-015-2692-4

Video saliency detection using 3D shearlet transform

Lei Bao [1] · **Xiongwei Zhang** [1] · **Yunfei Zheng** [2] · **Yang Li** [1]

Received: 26 August 2014 / Revised: 8 May 2015 / Accepted: 12 May 2015 /
Published online: 23 June 2015

Abstract Recently, visual saliency detection has received great interest. As most video saliency detection models are based on *spatiotemporal* mechanism, we firstly give a simple introduction of it in this paper. After discussing issues to be addressed, we present a novel framework for video saliency detection based on 3D discrete shearlet transform. Instead of measuring saliency by fusing spatial and temporal saliency maps, the proposed model regards video as three-dimensional data. By decomposing the video with 3D discrete shearlet transform and reconstructing it on multi-scales, this multi-scale saliency detection model obtains a number of feature blocks to describe the video. Based on each feature block, every a number of successive feature maps are taken as a whole, and the global contrast is calculated to obtain the saliency maps. By fusing all the saliency maps of different levels, the saliency map is generated for each video frame. This novel framework is very simple, and experimental results on ten videos show that the proposed model outperforms lots existing models.

Keywords Video saliency detection · 3D discrete shearlet transform · Feature blocks · Global probability density

1 Introduction

When viewing a scene, humans usually focus on some salient regions. This is a very useful mechanism working in our brain but hard to simulate with computer system. The area of saliency detection on static images or dynamic videos has been receiving more and more attention over the past few years (major achievements can be found in [2, 3]). Properly generated saliency maps can be useful for many applications, such as object location, detection, recognition, image or video compression. We will focus on video saliency detection in this paper.

✉ Lei Bao
lbaodong001@gmail.com

[1] College of Command Information Systems, PLA University of Science and Technology, Nanjing 210000, China

[2] Key Laboratory of Polarization Imaging and Detection, Officer Academy of Army, Hefei 230000, China

The saliency information of video exists not only on each separated frame but also between successive frames. In fact, the later one could be even more important in lots scenarios. Thus we cannot simply take image saliency of each frame as the final video saliency. To solve this problem, most existing video saliency detection models are based on *spatiotemporal* mechanism. They detect spatial saliency on single video frame and temporal saliency on inter-frame distinctiveness. The final saliency map is generated by fusing the spatial and temporal saliency maps together. Kim et al. [14] detected the spatial saliency with edge and color orientation information and the temporal saliency with absolute inter-frame distinctiveness, both based on center-surround framework. The final saliency is generated by linearly combining the spatial and temporal saliency with a fixed weight for each other. Instead of detecting saliency for all frames, Tapu et al. [24] detected saliency only for key frames extracted from tiny segments of the video. The spatial saliency is calculated with regional color information, the temporal saliency is calculated by detecting corresponding interest points between each key frame and its adjacent frames, and the final saliency is measured by using motion contrast to combine the spatial and temporal saliency. Li et al. [19] detected spatial saliency by computing color information of edge preserving super-pixels, which are extracted with advanced Turbopixels [18]. For temporal saliency, they used the same mechanism but on optical flow information of the video. The spatial and temporal saliency are then transmitted into conditional random field [17] to label each pixel. Li et al. [20] detected regional saliency for video frames. To segment each frame properly, the fast mean-shift process is performed on the spatial and temporal features computed from the color, texture and optical-flow information. The regional saliency is calculated by measuring the dissimilarity of each region with its neighbor regions. Afterward, these dynamic regions are matched in temporal domain to construct a temporally coherent regional saliency map. Rudoy et al. [23] detected video saliency with not only spatial and temporal information but also semantic information, and only computed saliency for some selected candidate gaze locations because saliency in video is very sparse according to their observation. Kim et al. [15] proposed an unified spatiotemporal saliency detection framework for both image and video based on textural contrast and motion stimuli. As this model focus on the contrast visual stimuli which can greatly eliminate unwanted details, it performs well even in complex scenes with highly textured backgrounds. Fang et al. [7] measured spatial saliency by extracting intensity, color, and texture features from DCT coefficients, then detected temporal saliency using motion feature in compressed domain, and designed a new fusion method to obtain the final saliency maps.

One main issue of *spatiotemporal* mechanism is computation redundancy. Most frames in a video have great relevance with their neighbors, thus saliency in video is very sparse [20]. Independent computing on every pixel of each frame is redundant. To address this issue, some researchers did not follow the above mentioned framework. Rapantzikos et al. [22] took video as a *spatiotemporal* volume, and detected video saliency by measuring local contrast for each visual unit. Duncan et al. [5, 6] built a model based on the theory of information entropy, that geometrically organized regions have lower entropy than disorganized regions. With Weighted Parzen Windows, they obtained the Renyi entropy of probabilistic relational distribution, based on distance and gradient direction relationships between pixels. This model emphasized biological plausibility in saliency detection process, while neglected the usefulness of color information.

Another even bigger issue of *spatiotemporal* mechanism is accuracy. As we know, human visual system is more sensitive to motion information than others. When building *spatiotemporal* saliency model, most researchers gave temporal saliency maps larger weight in the fusion process. But when the scene is static or has no significant motion, visual attention will be attracted by spatial

information, and spatial saliency should be given larger weight. The problem is that, no matter which saliency is more important, we actually have no idea what exactly the weight value should be. To solve this problem, researchers should use some proper mechanism to evaluate how powerful the motion contrast could be. Although it is a common sense that moving target draws people's attention, it could actually being wrong in some case. For example, leafs on a tree could be moving really fast in a windy day, but people usually still pay more attention to the beautiful bird on the trunk even if it is static. Or sometimes a static object draws more attention just because others are moving faster, like a walking man in the middle of a busy road. Thus no matter a fixed weight or dynamic weight is used to fuse the spatial and temporal saliency, the mechanism will still lose its accuracy due to the complexity of motion.

To overcome the above issues, we firstly give another point of view in video saliency detection – distinctiveness. Let's assume that people pay attention to distinctiveness while viewing a scene. This is also understandable as a common sense. When human visual system is attracted by moving object, we actually are attracted by the distinctiveness between frames, which is caused by the moving object. So does the distinctiveness in single frame, which is caused by color, shape or other features. That is to say, we can use distinctiveness to measure the saliency of each pixel. If the distinctiveness is larger, the saliency value should be higher and vise verse, no matter the distinctiveness is caused by spatial or temporal information.

In this paper, we propose a novel video saliency detection model, which regards the input video as three-dimensional data. Instead of using the input video straightly like Duncan et al. did in [5, 6], the video is firstly decomposed by 3D discrete shearlet transform to obtain multi-scale description. The reason of using shearlet based decomposition is to provide multi-scale analysis for saliency detection. The shearlet transform was originally introduced by Guo et al. [10]. It was derived within composite wavelet, which makes shearlets a truly multivariate extension of the wavelet framework. The use of shearing to control directional selectivity allows a single or finite set of generators to define shearlet systems. Although directional multi-scale systems have emerged years ago, only recently these representations have been extended beyond dimension 2. The extension of shearlet from 2-D to 3-D makes it possible for shearlet transform to analysis and process 3-D data sets like video. The 3-D shearlet representation is a multi-scale pyramid of well-localized wave-forms defined at various locations and orientations. It is introduced to overcome the limitations of traditional multi-scale systems while dealing with multi-dimension data. It has some good properties like parabolic scaling, directional sensitivity and spatial localization property. These are useful for saliency detection when describing the video frames in multi-scales and outstanding regions from their surroundings.

Instead of combining information of two dimensional spatial domain and one dimensional temporal domain, the proposed model is built on information of three-dimensional block. We actually need to deal with each video frame only once while at least twice for *spatiotemporal* mechanism. Most existing *spatiotemporal* saliency models make use of motion information between two frames, this would cause loss of long-term motion information. When viewing a scene, a number of successive frames would influence people's visual system, which is different from single images viewed independently. Taking this into consideration, we take one more step by processing the video per segment. Every frame is detected according to the feature maps of the segment it belongs to. As the proposed model could process more information, the detecting result could be more accurate as the experiments will show you later in this paper. What's more, it is no longer necessary to calculate meaningful weights as in fusing process of *spatiotemporal* mechanism.

2 The proposed saliency detection model

The proposed saliency detection model detects video saliency based on 3D discrete shearlet transform. Instead of using RGB color space for video saliency detection, all the video frames are converted to the Lab color space. Then each color channel of the converted video is decomposed with 3D discrete shearlet transform (3-D DShT) as presented in [21]. After denoising the obtained shearlet coefficients matrixes, feature blocks are generated by performing inverse shearlet transform on each decomposition level. On each feature block, global contrast is used to calculate saliency value. Then a saliency block is obtained for corresponding level. By fusing all the saliency blocks together, the final saliency value is calculated for each pixel. Thus we build the saliency map for each video frame. Figure 1 illustrates the overview of this novel saliency detection framework.

As shown in Fig. 1, there are mainly three steps to generate saliency maps for an input video. The first step is to convert all the video frames from RGB color space to Lab color space. In Lab color space, L is the intensity channel, a and b are RG and BY opponent channels respectively. The reason of this conversion is that Lab color space is more similar to human visual system, which would benefit the saliency evaluation mechanism.

The second step is to generate feature maps. First of all, all the converted video frames are resized to $m*m$. The target video is then segmented into a number of n-frames blocks. The last

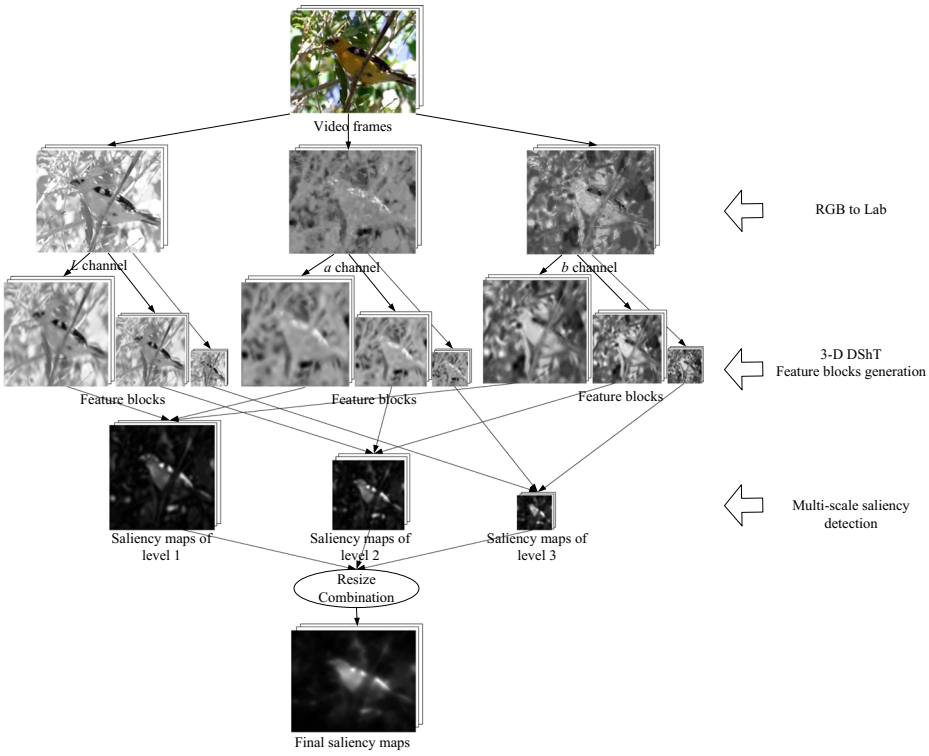


Fig. 1 Overview of the proposed saliency detection framework

frame of the last block will be copied to round up if needed. After that, by performing 3-D DShT on the resized video block v_0, the coefficients are obtained:

$$\left\{H_1^c, H_2^c, \ldots, H_M^c, L^c\right\} = SH(v_0{}^c) \tag{1}$$

where c represents the channel of input video block v_0 as $c \in \{L,a,b\}$, H_i^c represents the shearlet coefficient matrix of the i-th level on channel c, M represents the maximum decomposition level, L^c represents the scaling coefficients matrix on channel c, SH represents the 3-D DShT process. The coefficient matrixes of different levels have different sizes. For example, when m=192, n=96, M=3, the size of H_1^c is 192∗192∗96, H_2^c is 128∗128∗64, H_3^c is 64∗64∗32, L^c is 32∗32∗16.

The shearlet coefficients corresponding to salient regions are larger than others. For de-noising, we remove non-salient coefficients by setting all shearlet coefficients less than a proper threshold to be 0. And enlarge saliency by setting other coefficients to be w times larger. The formula is:

$$H'_{i,j} = \begin{cases} wH_{i,j} & \left|H_{i,j}\right| \geq \delta \\ 0 & \left|H_{i,j}\right| < \delta \end{cases} \tag{2}$$

where δ is the threshold obtained by Visu Shrink Threshold Function [4].

The shearlet coefficients matrix on each level represents the detailed information of the video at the corresponding level. And the scaling coefficients matrix represents the approximation information at the coarsest resolution. To create the l-th feature block, we perform inverse shearlet transform on all the shearlet coefficients of lower levels (H_1', H_2',…,H_l') and the scaling coefficients L, as to be shown in (3).

$$f_l^c = ISH\left(H'^c_1, H'^c_1, \ldots, H'^c_l, L^c\right) \tag{3}$$

where ISH represents the inverse discrete shearlet transform process, l represents the decomposition level. Equation (3) creates M feature blocks, each for a decomposition level. And the l-th feature block f_l^c is the same size as $H_l^{c\prime}$.

The third step is to calculate saliency value on each decomposition level, and generate saliency map for each video frame. In this paper, we detect video saliency by measuring the rarity of different regions with global contrast. When we detect image saliency in our former paper [1], the saliency is measured in two aspects: global and local. In this paper, we regards the input video as three-dimensional data. If we use local contrast to measure video saliency, all the distinctiveness would be magnified including the distinctiveness exist in single frame's background or caused by background movements, thus the non-salient background would be labeled with large saliency values. That is the reason why we only use global contrast to detect the video saliency in this paper.

As we know, video saliency is evaluated by measuring contrast based on features like color, luminance, texture and motion. Figure 2 shows some examples. It can be seen that Fig. 2a can use color information to outstand region P, Fig. 2b can use luminance information to outstand region P, Fig. 2c can use direction information to outstand region P, Fig. 2d can use shape information to outstand region P. No matter which feature works, it can be concluded that the rarity draws attention. Taking Fig. 2a as an example, the reason of P outstanding from its surroundings is that the average color of region P is green, while other regions is red. That is to say, in this example, the smaller area covered by green color draws more attention than larger area covered by red color. Of course, this conclusion is valid if and only if there is no personal reference. Different from the former four examples, Fig. 2e is more complicated. Neither the left darker regions nor the right brighter regions, but the light and dark alternate regions are salient. That is to say, we can't simply use luminance to outstand region P. Using rarity to explain, as only the alternate region covers both light and dark area

(see region labeled by red box), the region P is salient. That is to say, the rarity of the alternate region attracts our attention.

The mainly difference of video from image is inter-frame motion information. In temporal, whether a region is outstanding from its surroundings depends on whether it shows different motion. If there exist only one salient object in the input video, there are mainly two kinds of motion styles. One is the global motion caused by the background (no matter the background is moving or not), the other is the motion caused by the salient object. In general, the region covered by moving salient object is usually much smaller than the remainder. This conclusion is still valid if there exist more salient objects.

Before calculating the saliency values, we need to define every location. By using all the feature blocks, every location (x,y,t) on the l-th level can be represented by a feature vector $f_l(x,y,t)$ as:

$$f_l(x,y,t) = \left[f_l^L(x,y,t), f_l^a(x,y,t), f_l^b(x,y,t)\right]^T \tag{4}$$

Then global contrast is used to generate saliency maps for video frames on each decomposition level. As mentioned above, the saliency of the t-th video frame vf_t is not only affected by itself, but also a number of neighbor frames. To be simple, every h successive frames are processed as a whole (h=4 in this work), and we set $\varepsilon=(vf_t, vf_{t+1}, \ldots, vf_{t+h-1})$. For every location $(x,y,t)\in\varepsilon$, the likelihood of the features can be defined as the probability density handled by a normal distribution as:

$$p_l(x,y,t) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \times e^{-\frac{1}{2}(f_l(x,y,t)-\mu)^T \Sigma^{-1}(f_l(x,y,t)-\mu)} \tag{5}$$

where $\Sigma=E[(f_l(x,y,t)-\mu)(f_l(x,y,t)-\mu)^T]$ is the covariance matrix, μ represents the expectation vector, $f_{l,t}$ represents the feature map on the l-th level for the t-th video frame. The saliency value of location (x,y,t) on the l-th level is defined as:

$$S_l(x,y,t) = G(-\log_{10} p_l(x,y,t)) * I_{k*k} \tag{6}$$

where I_{k*k} represents a 2-D Gaussian low-pass filter (k=5 in this work), which is employed to get a smoother result. $G(\cdot)$ represents used to convert $(\cdot)$ to a grayscale image. It is worth to mention that some of the feature blocks may equal to zero matrix and the determinant of their covariance matrix is zero. This would cause the probability density defined in (5) to be invalid. Thus only the non-zero feature blocks are used to generate feature vectors.

As the size of S_l obtained by (6) vary between different levels, the size of saliency maps is interpolated to be the same as first level, so does the quantity of saliency maps. For the t-th video frame, by fusing all levels' saliency maps together, we obtain:

$$S(t) = \sum_{l=1}^{M} N(S_l(t)) \tag{7}$$

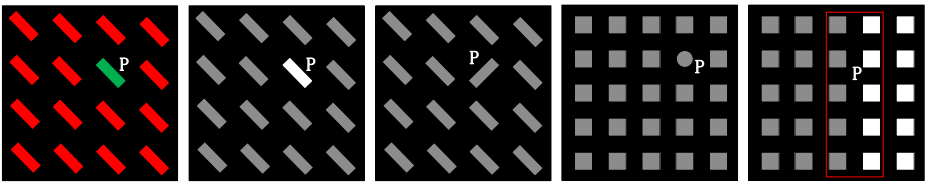


Fig. 2 Examples of visual saliency

where $N(\cdot)$ represents the normalization operator [13], $S_l(t)$ represents the saliency map on the l-th level for the t-th video frame.

Goferman et al. [9] pointed out that the location around the focus of attention point should be given larger saliency value than those far away from it. Here, the locations with saliency value larger than 0.8 as in [9] are signed as focus of attention points. The final saliency map is calculated as

$$S_0(x,y,t) = S(x,y,t) \times (1-d(x,y,t)) \tag{8}$$

where $d(x,y,t) = {}^{d_0(x,y,t)}/_{\sqrt{a_0{}^2+b_0{}^2}}$, $d_0(x,y)$ represents the distance between location (x,y) and the nearest focus of attention point (x_0,y_0), $a_0 * b_0$ is the size of $S(t)$.

In order to understand how the proposed model can be applied in practice as well as for the reconstruction of the experiments for validation purposes, follows are the overall flowchart and detailed pseudo-code (Fig. 3).

Algorithm: Pseudo-code for proposed video saliency detection model

Input: Video frames v

Generate L, a, b channels' images v^L, v^a, v^b.
for each channel $c \in \{L,a,b\}$
  Obtain shearlet coefficient matrix H_i^c and scaling coefficients matrix L^c by 3-D DShT.
  for each shearlet coefficient matrix H_i^c
    if $|H_{i,j}| \geq \delta$
      $H'_{i,j} = wH_{i,j}$
    else
      $H'_{i,j} = 0$
    end
  end
  Generate feature blocks $f_l^c = ISH(H_1'^c, H_1'^c, ..., H_l'^c, L^c)$.
end
for each decomposition level $l \in \{1,...,M\}$
  Generate feature vector $f_l(x,y,t) = [f_l^L(x,y,t), f_l^a(x,y,t), f_l^b(x,y,t)]^T$.
  for every h successive frames $\varepsilon = (vf_t, vf_{t+1}, ..., vf_{t+h-1})$
    $$p_l(x,y,t) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \times e^{-\frac{1}{2}(f_l(x,y,t)-\mu)^T \Sigma^{-1}(f_l(x,y,t)-\mu)}.$$
    $S_l(x,y,t) = G(-\log_{10} p_l(x,y,t)) * I_{k*k}$.
  end
end
Generate saliency maps by fusing different levels' salient values: $S(t) = \sum_{l=1}^{M} N(S_l(t))$.
Generate the final saliency maps by updating: $S_0(x,y,t)$ $S(x,y,t)$ (1 $d(x,y,t)$).

Output: Saliency maps S_0

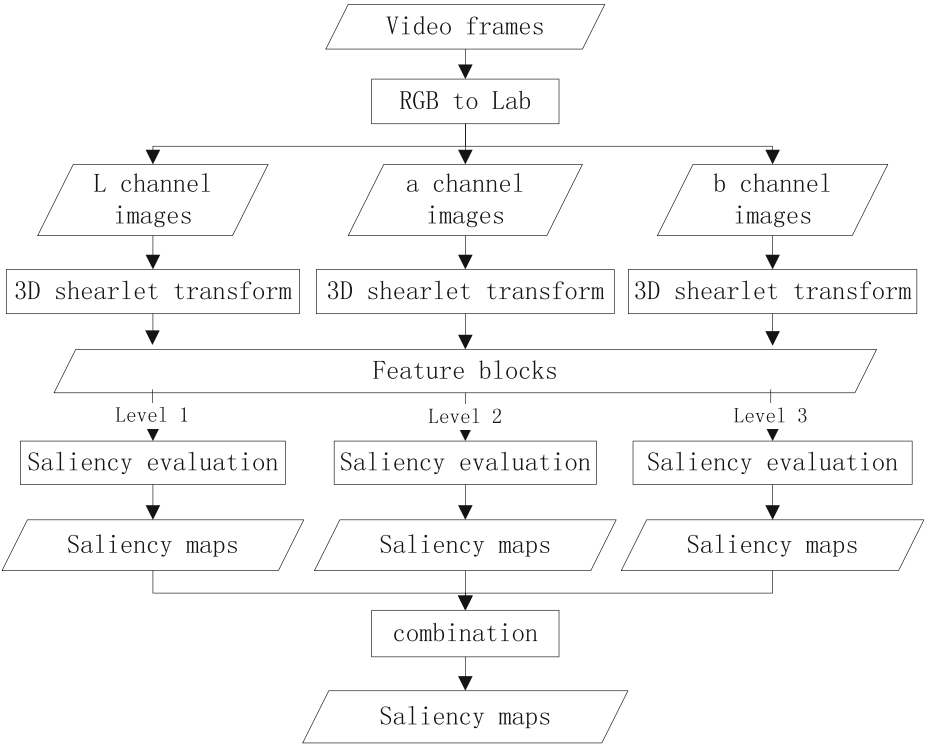


Fig. 3 Flowchart of the proposed video saliency detection model

3 Experiments and evaluations

In this section, we evaluate the performance of the proposed model with database from the website of Akisato [8]. This database includes ten videos, named AN119T, BR128T, BR130T, DO01_013, DO01_014, DO01_030, DO01_055, DO02_001, M07_058 and VWC102T. To be simple, V1-V10 are used to represent these ten videos respectively. V1-V10 cover situations with different complexities. For example, V4 and V9 have very pure backgrounds, both are clear blue sky. The background of V5, V7 and V10 are less pure but the global movement can be easily compensated. Part of V8's background moves violently. It may confuse the detection process when detecting temporal saliency in *spatiotemporal* framework. V2 and V3 have the most complex moving background, which is difficult for saliency detection [20, 25]. Each video includes about 100 frames and its corresponding ground-truths (see Fig. 4b). Instead of using the original ground-truths, they are firstly converted to binary maps (see Fig. 4c). In the ground-truths, we set pixels of salient regions to be 1, pixels of non-salient regions to be 0, and then the binary maps are built. Fig. 4d shows some saliency detection result of [20]. By comparing with the results of the proposed model (see Fig. 4e), we can find that although the proposed model detects lower saliency on most of the videos, but it gives a cleaner saliency maps since it is more robust to noise. In another word, much less non-salient regions are signed to be salient comparing with the model of [20].

Itti et al. proposed a bottom-up model in 1998 [14]. It is the first model who completely implement and verify Koch and Ullman model [16], and the most classical one who has great influence on the whole visual saliency detection area. By merging motion and flicker feature

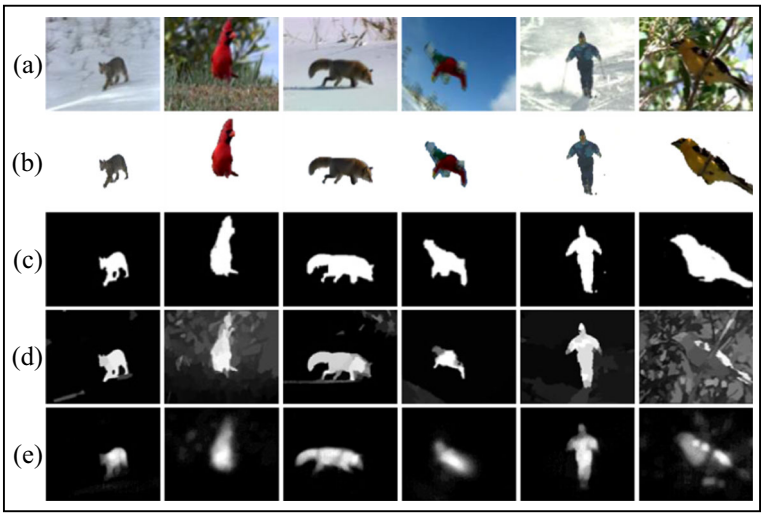


Fig. 4 Examples of different approaches

channels in Itti's model, Harel further presented an implementation of video saliency detection [12]. We will compare the proposed model with Harel's work referred as HAREL. Furthermore, Duncan et al.'s model RE in [6], Zhou et al.'s model TM in [26], Hadizadeh et al.'s model SAVC in [11] are also taken in comparison. These models cover several categories of video saliency detection mechanism. RE regards video as three-dimensional data, and introduces information entropy theory. TM and SAVC are state-of-the-art models. Both of them are based on *spatiotemporal* framework, while SAVC estimates saliency in the DCT domain based on Itti-Koch-Niebur saliency model [8]. Some of the saliency detection results are shown in Figs. 5 and 6, while (a) represents original video frames, (b) represents results of HAREL, (c) represents results of TM, (d) represents results of SAVC, (e) represents results of RE, (f) represents results of the proposed model. For HAREL, TM, SAVC and the proposed model, we show the 10th, 20th, 30th, 40th, 50th, 60th frames' saliency maps. As each saliency map generated by [6] corresponds to m successive frames (m=5 in [6]), we take the 2th, 4th, 6th, 8th, 10th, and 12th saliency map for comparing. Take the 4th saliency map as an example, this saliency map is generated on 16th to 20th video frames, thus the 4th saliency map can be taken as these 5 frames' saliency map when comparing with other models.

By comparing the saliency results of different models stated in Figs. 5 and 6, it is obvious that the proposed model outperforms other four models, especially with complex background (see BR128T and BR130T). From Figs. 5 and 6, we can find that HAREL is more sensitive to noise, and would cause more irrelevant salient regions than the proposed model (see DO01_055 and VWC102T). Besides, HAREL detects saliency based on center-surround mechanism. It would lose the regions inside salient objects, or give lower saliency value for these inside pixels (see AN119T and DO01_014). SAVC and TM perform better on detecting saliency inside salient objects. But these two models are even more sensitive to noise than HAREL, also much more non-salient regions are signed to be salient than the proposed model. For example, the background of DO01_055, DO01_030 and M07_058 is clean sky, which is obvious non-salient. But SAVC and TM still signed lots massive salient regions in these areas. RE is more robust to noise than other three models. But RE is built on local information

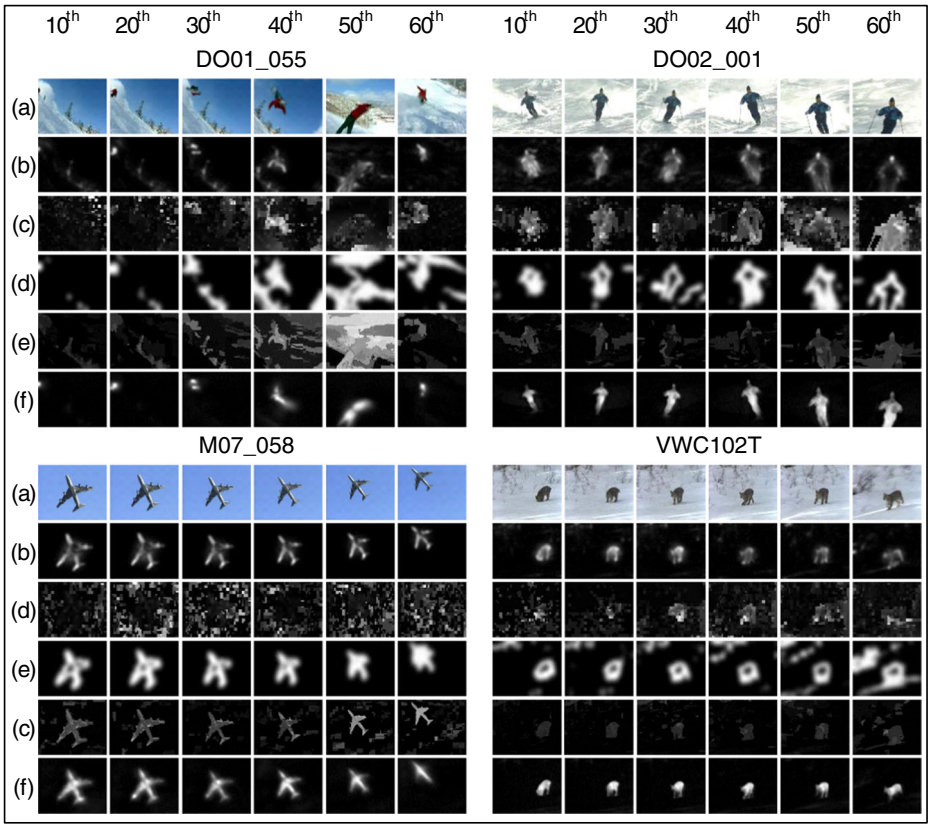


Fig. 5 Examples of different approaches

entropy, which would lose salient regions inside salient objects as HAREL. Besides, RE is inclined to sign non-salient regions around salient objects salient. It would generate more irrelevant salient regions than the proposed model.

Beside comparing different models directly on saliency maps, we further evaluate the performance of the proposed model based on *Precision* (P for short), *Recall* (R for short) and F_a *measure* (F_a for short) values by using the ground-truths. The definitions of P, R and F_a are as follows:

$$P = \frac{\sum_{i=1}^{N} P_i}{N} \quad as \quad P_i = \frac{sum(gt*s)}{sum(s)} \tag{9}$$

$$R = \frac{\sum_{i=1}^{N} R_i}{N} \quad as \quad R_i = \frac{sum(gt*s)}{sum(gt)} \tag{10}$$

$$F_a = \frac{(1+a) \times P \times R}{a \times P + R} \tag{11}$$

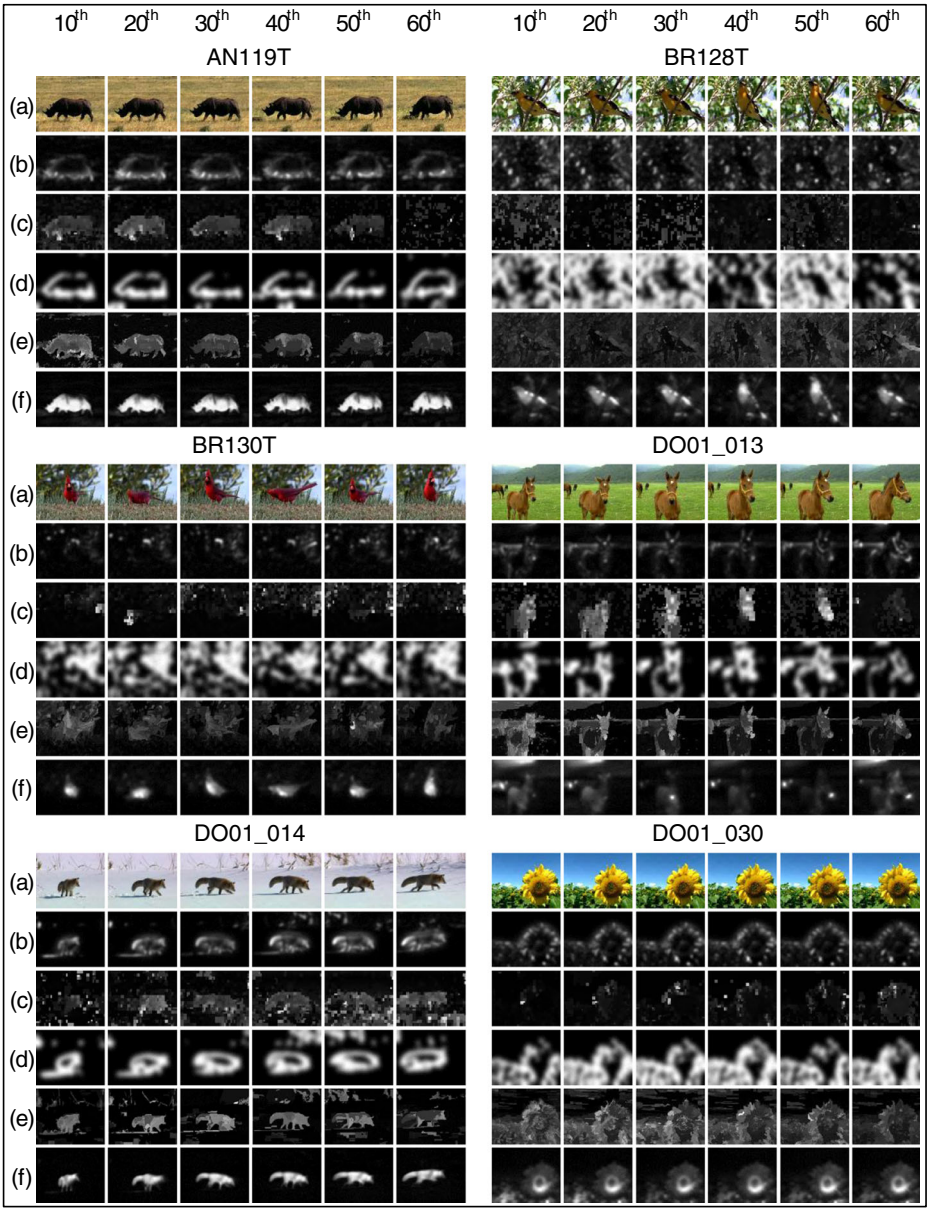


Fig. 6 Examples of different approaches

where *gt* represents the binary map, *s* represents saliency map, N is the quantity of images, a is chosen to be 0.3 as most saliency detection models do. Since some of the video frames have not been given ground-truths in the database, we only use the video frames with ground-truths to obtain P, R and F_a values (starting from the 16th video frame).

The overall P, R and F_a results are shown in Fig. 7, while (a) represents P, R and F_a values obtained by using mean value as threshold to generate binary maps, (b) represents P, R and F_a

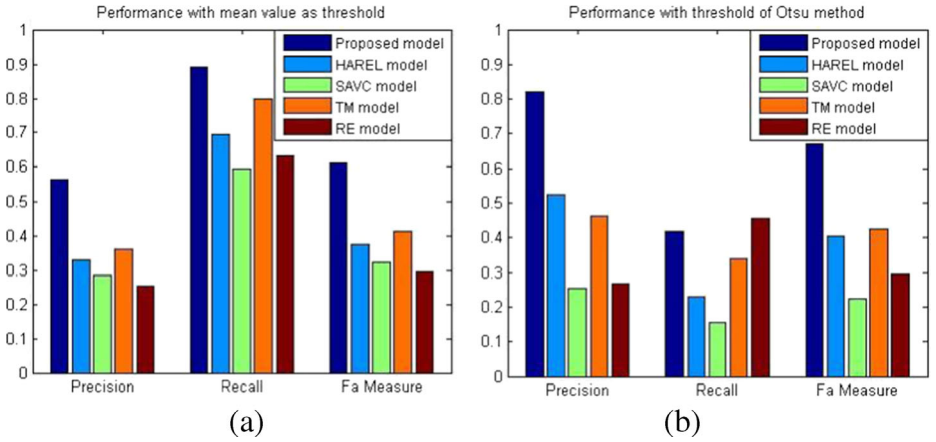


Fig. 7 Performance comparison with P, R and F_a values

values obtained by using Otsu function to generate binary maps. From Fig. 7, we can find that the overall performance of the proposed model is better than HAREL, SAVC and RE. By comparing the proposed model with TM, when we use mean value as threshold, the proposed model performs better than TM. When we use the Otsu function to obtain threshold, the proposed model has higher P and F_a values, but lower R values than TM. One of the main reasons is that TM generates more irrelevant salient regions than the proposed model, which would cause higher R values with lower P values. Another reason is that when using Otsu function to generate threshold, the binary maps of the propose model would lose some salient regions, which would cause higher P values with lower R values.

Figure 8 shows binary maps obtained by different thresholds on saliency maps of the proposed model, while (a) represents input video frames, (b) represents saliency maps of the proposed model, (c) represents ground-truths, (d)-(k) represent binary maps of different thresholds $T=8*k(k=\{1,2,\ldots,8\})$. From Figure 8, we can find that the regions covered by positive value become smaller when the threshold becomes larger. That is to say, the method of setting threshold would influence performance evaluation results, if the evaluation method is built on binary maps. And this is why Fig. 7a and b obtain different comparison results.

To have a better comparison, we use different thresholds (0–255) to obtain binary maps. After calculating the P, R and F_a values, we draw the PR curve, which can be seen in Fig. 10a. Figure 10b shows the average P, R and F_a values. From Fig. 10a, we can see that the PR curve of the proposed model is closer to the (1,1) point, which means the proposed model outperforms other four models. From Fig. 10b, we can see that the proposed model performs better than HAREL, SAVC and TM. Comparing with RE, the proposed model has obvious higher P and F_a values, while R values is a little lower than RE. The reason is that RE generates more irrelevant salient regions (see Fig. 9), which can obtain higher R values with lower P values. It can be concluded that the proposed model outperforms RE (Fig. 10).

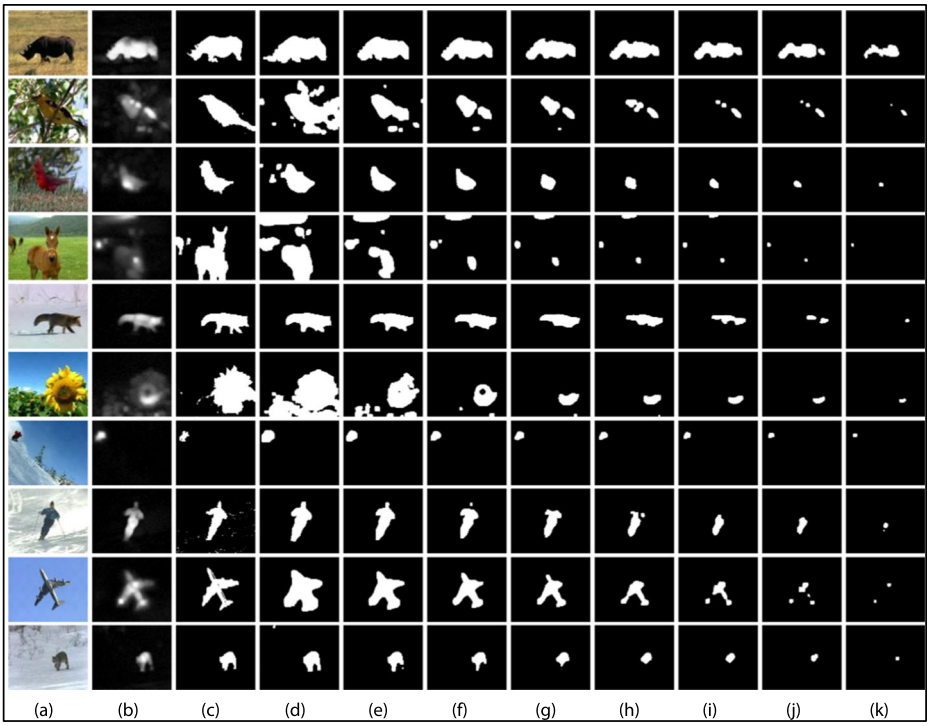


Fig. 8 Binary maps obtained by using different thresholds of the proposed model

As the ten videos cover situations on different complexities, we further present the PR curves and average P, R, F_a values for different saliency detection models on each video, which can be seen in Fig. 11. From PR curves, we can see that the proposed model performs better on most of the videos, especially on BR128T, BR130T and DO01_014. From the average P, R and F_a values, we can see that the proposed model has higher P, R and F_a values than the other models on AN119T, BR128T, BR130T, DO01_014 and DO02_001, but lower P, R and F_a values on DO01_013. Besides, the proposed model has higher P and F_a values, lower R value than other models on DO01_030, DO01_055 and VWC102T; lower P values, higher R and F_a values than HAREL, higher P and F_a values, lower R values than RE on M01_058. According to Fig. 11, on one hand, the PR curves of the proposed model is closer to (1,1) point on DO01_030, DO01_055, VWC102T and M07_058. On the other hand, the proposed model has higher F_a values on these four videos. It can be concluded that the proposed model performs better on DO01_030, DO01_055, VWC102T and M07_058.

From Fig. 11, we can see that the proposed model performs worse on DO01_013. The reason is that the salient objects in DO01_013 cannot be well distinguished from the background in L, a and b channels. But the proposed model is built on color information of L, a and b channels. Improving the proposed model to perform better on videos like DO01_013 would be one of our further works.

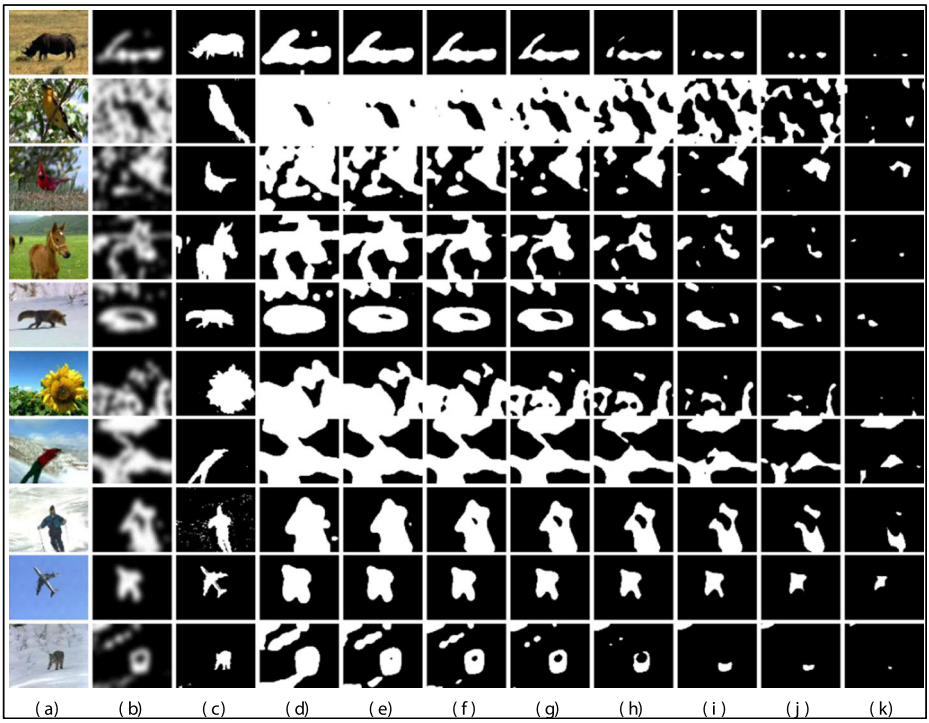


Fig. 9 Binary maps obtained by using different thresholds of RE

4 Conclusions

In this paper, a novel video saliency detection framework using 3-D DShT is proposed. It begins with generating feature blocks in multi-scales by performing inverse shearlet transform. Saliency blocks are then calculated accordingly. The final detecting result is a combination of saliencies on different levels. Comparing with the popular *spatiotemporal* mechanism, our framework takes video as 3D information and

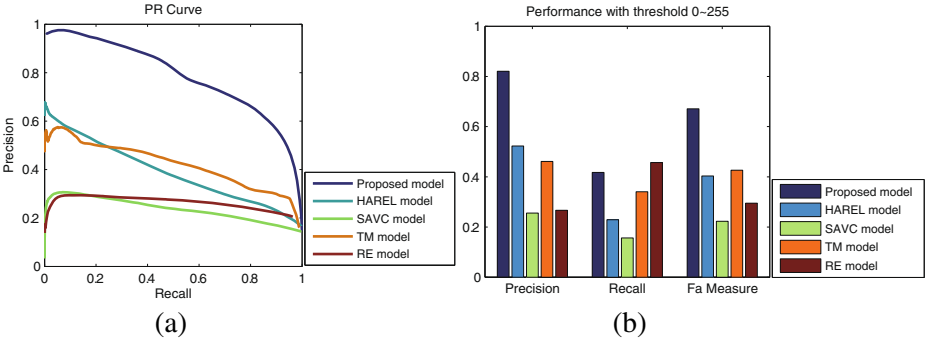


Fig. 10 PR curves and average P, R and F_a values for different saliency detection models

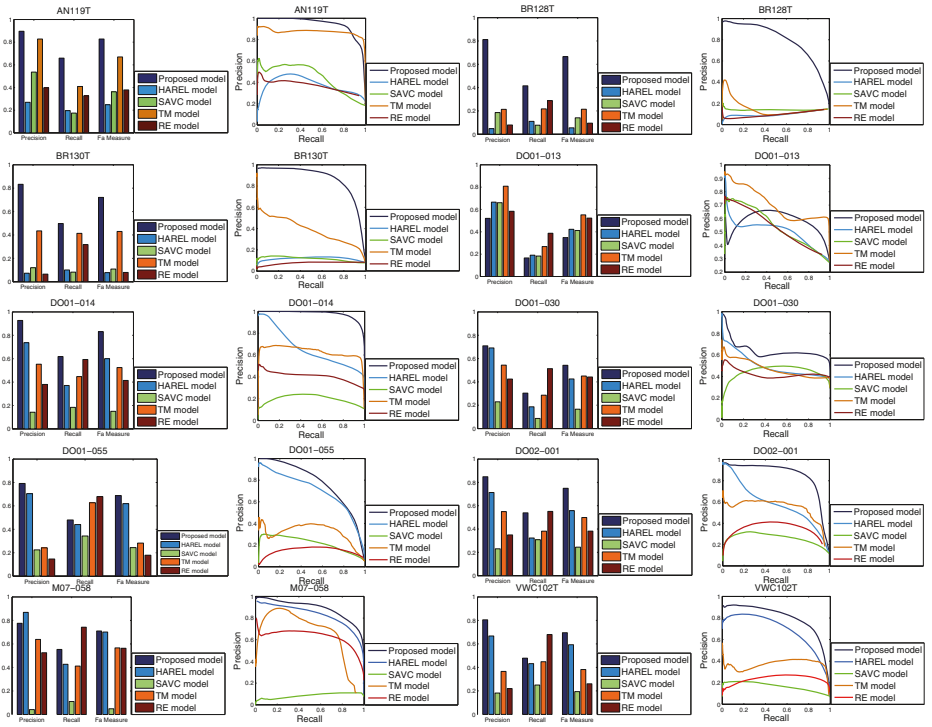


Fig. 11 PR curves and average P, R and F_a values for different saliency detection models on different videos

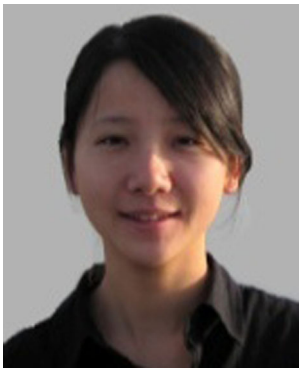
stands on one exclusive detecting base - distinctiveness. To the best of our knowledge, the work in this paper is the first try to detect video saliency regions on shearlet domain, and the experimental results demonstrate the performance of the new proposed model.

The proposed framework is extendable. In the future, we will further explore how to improve the performance by combining texture, direction and other features. Also, the employ of 3-D DShT requires to load every a number of successive video frames in memory for processing. This would limit the use of the proposed framework in some real-time applications.

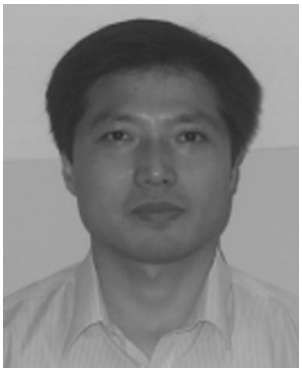
References

1. Bao L, Lu J, Li Y, Shi Y (2015) A saliency detection model using shearlet transform. Multimedia Tools Appl 74(11)
2. Borji A, Itti L (2013) State-of-the-art in visual attention modeling. IEEE Trans Pattern Anal Mach Intell 35(1):185–207
3. Borji A, Sihite DN, Itti L (2013) Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. IEEE Trans Image Process 22(1):55–69
4. Donoho DL (1995) De-noising by soft-thresholding. IEEE Trans Inf Theory 41(3):613–627
5. Duncan K, Sarkar S (2010) REM: relational entropy-based measure of saliency. Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing, pp. 40–47

6. Duncan K, Sarkar S (2012) Relational entropy-based saliency detection in images and videos. Image Processing, 2012 19th IEEE International Conference on, pp. 1093–1096
7. Fang Y, Lin W, Chen Z, Tsai C, Lin C (2013) A video saliency detection model in compressed domain. IEEE Trans Circuits Syst Video Technol 24(1):27–38
8. Fukuchi K, Miyazato K, Kimura A, Takagi S, Yamato J (2009) Saliency-based video segmentation with graph cuts and sequentially updated priors. IEEE Int Conf Multimedia Expo 638–641
9. Goferman S, Zelnik-Manor L, Tal A (2012) Context-aware saliency detection. IEEE Trans Pattern Anal Mach Intell 34(10):1915–1926
10. Guo K, Kutyniok G, Labate D (2006) Sparse multidimensional representations using anisotropic dilation and shear operators. Wavelets Splines 189–201
11. Hadizadeh H, Bajic IV (2014) Saliency-aware video compression. IEEE Trans Image Process 23(1):19–33
12. Harel J, Koch C, Perona P (2006) Graph-based visual saliency. Proc Neural Inf Process Syst 545–552
13. Itti L (2000) Models of bottom-up and top-down visual attention. California Institute of Technology
14. Kim W, Jung C, Kim C (2011) Spatiotemporal saliency detection and its applications in static and dynamic scenes. IEEE Trans Circuits Syst Video Technol 21(4):446–456
15. Kim W, Kim C (2013) Spatiotemporal saliency detection using textural contrast and its applications. IEEE Trans Circuits Syst Video Technol 24(4):646–659
16. Koch C, Ullman S (1987) Shifts in selective visual attention: towards the underlying neural circuitry. Matters of intelligence, ed: Springer, pp. 115–141
17. Lafferty J, McCallum A, Pereira FC (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning, pp. 282–289
18. Levinshtein A, Stere A, Kutulakos KN, Fleet DJ, Dickinson SJ, Siddiqi K (2009) Turbopixels: fast superpixels using geometric flows. IEEE Trans Pattern Anal Mach Intell 31(12):2290–2297
19. Li W, Chang H, Lien K, Chang H, Wang Y (2013) Exploring visual and motion saliency for automatic video object extraction. IEEE Trans Image Process 22(7):2600–2610
20. Li Y, Sheng B, Ma L, Wu W, Xie Z (2013) Temporally coherent video saliency using regional dynamic contrast. IEEE Trans Circuits Syst Video Technol 23(12):2067–2076
21. Negi PS, Labate D (2012) 3-D discrete shearlet transform and video processing. IEEE Trans Image Process 21(6):2944–2954
22. Rapantzikos K, Tsapatsoulis N, Avrithis Y, Kollias S (2009) Spatiotemporal saliency for video classification. Signal Process Image Commun 24(7):557–571
23. Rudoy D, Goldman DB, Shechtman E, Zelnik-Manor L (2013). Learning video saliency from human gaze using candidate selection. Computer Vision and Pattern Recognition, 2013 I.E. Conference on, pp. 1147–1154
24. Tapu R, Zaharia T (2012) Video structuring: from pixels to visual entities. Signal Processing Conference, Proceedings of the 20th European, pp. 1583–1587
25. Wu B, Xu L, Zeng L, Wang Z, Wang Y (2013) A unified framework for spatiotemporal salient region detection. EURASIP J Image Video Process 2013(1):1–12
26. Zhou F, Kang S, Cohen M (2014) Time-mappingusingspace-timesaliency. Computer Vision and Pattern Recognition, 2014 I.E. Conference on, pp. 3358–3365



Lei Bao received her M.S. degrees in computer science and technology from Electronic Engineering Institute of PLA, Hefei, China, 2011. She is currently pursuing the Ph.D. degree in computer science and technology from PLA University of Science and Technology. Her research interests include image/video processing, and machine learning.



Xiongwei Zhang received his Ph.D. degree in computer science from PLA University of Science and Technology, Nanjing, China, 1992. Now, he is a Professor at PLA University of Science and Technology. His research interests include multimedia information processing, digital communication, computational intelligence.

Yunfei Zheng received his M.S. degrees in signal and information processing from Artillery Academy of PLA, 2008. He is currently pursuing the Ph.D. degree in signal and information system from PLA University of Science and Technology. Her research interests include image video processing and machine learning.

Yang Li received his M.S. degree in computer science and technology from PLA University of Science and Technology, Nanjing, China, 2010. Now, he is an Assistant Professor at PLA University of Science and Technology. His research interests include digital signal and image processing.