SALIENT OBJECT DETECTION WITH SPATIOTEMPORAL BACKGROUND PRIORS FOR VIDEO

[#1]SYED TAMEEM BASHA QUADRI, M.TECH STUDENT
[#2]MR.T.VIJAY KUMAR, ASSOCIATE PROFESSOR
[#1,2]DEPT OF ECE
[#1,2]*Dr.K.V.SUBBA REDDY INSTITUTE OF TECHNOLOGY , KURNOOL*

ABSTRACT

Saliency detection for images has been studied for many years, for which a lot of methods have been designed. In saliency detection, background priors which are often regarded as pseudo-background are effective clues to find salient objects in images. Although image boundary is commonly used background priors, it doesn't work well for images of complex scenes and videos. In this paper, we explore how to identify the background priors for a video and propose a saliency based method to detect the visual objects by using background priors. For a video, we integrate multiple pairs of SIFT flows from long-range frames and a bidirectional consistency propagation is conducted to obtain the accurate and sufficient temporal background priors, which are combined with spatial background priors to generate spatiotemporal background priors. Next, a novel dual-graph based structure using spatiotemporal background priors is put forward in computation of saliency maps, fully taking advantage of appearance and motion information in videos. Experimental results on different challenging datasets show that the proposed method robustly and accurately detect the video objects in both simple and complex scenes and achieve better performance compared with other state-of-the-art video saliency models.

I.INTRODUCTION

Salient object detection has been a very popular topic in recent years, which is evolved from visual attention research for images and videos. Different from traditional saliency research which refers to the prediction of eye fixations, salient object detection aims to find visual objects by measuring the dissimilarity of each region based on the fact that there exists statistic divergence between foreground region and background region. In application, salient object detection can be utilized in plenty of fields, such as image segmentation, image localization, image compression and so on. Particularly, this paper focuses on salient object detection.

Many methodologies have been proposed to detect objects in an image or video. Instead of the traditional center-surrounding weight mechanism, more and more methods employ the concept "background prior", which can be regarded as a set of cues or templates for the background and shows a higher performance against conventional salient object detection methodologies. At present, most of the published papers treat the image boundary as background priors based on the empirical conclusion that for most images, the salient object will not appear on the boundary. However, this is not always the case: in certain circumstances, the image boundary is vulnerable for background priors.

1.1 Introduction to salient object detection from videos

Salient object detection from videos plays an important role as a pre-processing step in many computer vision applications such as video re-targeting, object detection, person reidentication and visual tracking. Conventional methods for salient object detection often segment each frame into regions and artificially combine low-level (bottom-up) features (e.g., intensity , color , edge orientation) with heuristic (top-down) priors (e.g., center prior , boundary prior , abjectness) detected from the regions. Low-level features and priors used in the conventional methods are hand-crafted and are not sufficiently robust for challenging cases, especially when the salient object is presented in a low-contrast and cluttered background. Although machine learning based methods have been recently developed , they are primary for integrating different hand-crafted features or fusing multiple saliency maps generated from various methods. Accordingly, they usually fail to preserve object details when the salient object intersects with the image boundary.

Fig.1.: Top row images are original video frames, followed by the ground truth and corresponding saliency maps obtained below.

Has similar appearance with the background where hand-crafted features are often unstable. Recent advances in deep learning using Deep Neural Network (DNN) enable us to extract visual features, called deep features, directly from raw images/videos. They are more

powerful for discrimination and, furthermore, more robust than hand-crafted features. Indeed, saliency models for videos using deep features have demonstrated superior results over existing works utilizing only hand-crafted features. However, they extract deep features from each frame independently and employ frame-by-frame processing to compute saliency maps, leading to inaccuracy for dynamically moving objects. This is because temporal information over frames is not taken into account in computing either deep features or saliency maps. Incorporating temporal information in such computations should lead to better performance. Computed saliency maps do not always accurately reflect the shapes of salient objects in videos. To segment salient objects as accurately as possible while reducing noise, dense Conditional Random Field (CRF), a powerful graphical model to globally capture the contextual information, has been applied to the computed saliency maps, which results in improvement in spatial coherence and contour localization. However, dense CRF is applied to each frame of a video separately, meaning that only spatial contextual information is considered. Again, temporal information over frames should be taken into account for better performance.

Motivated by the above observation, we propose a novel framework using spatio temporal information as fully as possible for salient object detection in videos. We introduce anew set of Spatio Temporal Deep (STD) features that utilize both local and global contexts over frames. Our STD features consist of local and global features. The local feature is computed by aggregating over frames deep features, which are extracted from each frame using a region-based Convolution Neural Network (CNN). The global feature is computed
from a temporal-segment of a video using a block-based. We also introduce the Spatio Temporal CRF (STCRF), in which the spatial relationship between regions in a frameas well as temporal consistency of regions over frames is formally described using STD features. Our proposed method first segments an input video into multi-scale levels, and thenat each scale level, extracts STD features and computes a saliency map. The method then fuses saliency maps at different scale levels into the final saliency map. Extensive experiments on public benchmark datasets for video saliency confirm that our proposed method significantly outperforms the state-of the-arts. Example of saliency maps obtains by our method. We also apply our method to video object segmentation and observe that our method outperforms existing methods.

II. LITERATURE SURVEY

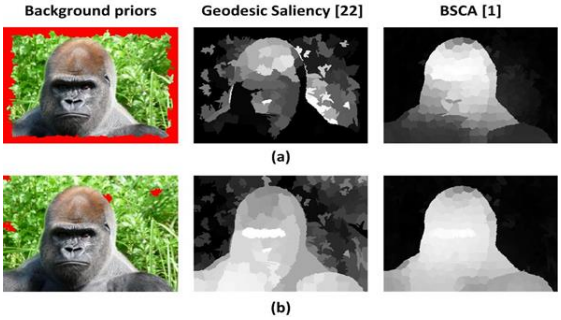


Fig.2: What are good background priors?

As can be seen in Fig 2, red regions of the first image represent the image boundary as background priors. The second and third images are saliency maps generated respectively by two state-of-the-art salient object detection. Approaches using image boundary as background priors, From these two images, we can see that the salient object cannot be completely detected and also the background is mistakenly regarded as components of the salient object. However, in Fig 2.1(b), we manually select a few patches from the background as background priors, which highlight uniform detection results from the background.

So what can be called good background priors? As Fig 2.1 shows, good background priors possess the following two characteristics:

1)As accurate as possible. In other words, background priors should contain minimum or even no contents of the foreground.

2) As sufficient as possible. That means background priors ought to include as much background as we can.

The most relevant field to our topic is video object segmentation via background subtraction. Different from background subtractions, background priors don't need to cover all back-ground regions. Seen from the above analysis, the existing methods still have certain space for further improvement because of their respective defects.

III. EXISTING SYSTEM

In this section, we assess the approach in saliency segmentation experiments. The performance is compared with the state-of-the-art methods using the programs given by the authors or our own implementation with default parameters. The experiments are divided into two parts, where the first one considers still images and the second one video sequences. Existing system have low performance and it taken by object detection using video segments.

3.1 Segmenting salient objects from images

First, we run the publicly available saliency segmentation test, introduce in the method is compared to the band-pass which was reported to

achieve clearly the best performance among the several tested methods

The experiment contains 1000 color images with pixel-wise ground truth segmentations provided by human observers. First a saliency map is computed for each test image and then segmentation is generated by simply thresholding the map by assigning the pixels above the given threshold as salient (white foreground) and below the threshold as non-salient (black background). A precision and recall rate is then computed using definitions:

$$precision = |SF \cap GF|/|SF|, \qquad recall = |SF$$

where SF denotes the segmented foreground pixels, GF denotes the ground truth foreground pixels, and $|\cdot|$ refers to number of elements in a set. By sliding the threshold from minimum to maximum saliency value, we achieved the precision recall curves illustrated in Figure 3 (magenta, cyan, orange, and green).

The results show that the proposed saliency measure achieves the highest performance up to a recall rate 0.9. Furthermore also the method from seems to outperform the state-of-the-art result. Notice that the precision recall curves of the proposed method and the method do not have values for small recalls because several pixels reach the maximum saliency value and they change labels simultaneously when the threshold is lowered below one. At maximum recall all methods converge to 0.2 precision, which corresponds to a situation where all pixels are labeled as foreground.

We continue the experiment by adding the CRF segmentation model from Section 3 on top of our saliency measure. First, we perform the same experiment as above, but refine the threshold saliency maps using the CRF model (i.e. the first choice is used for f). The resulting precision-recall-curve in Figure 3 (blue) illustrates a clear gain compared to threshold saliency map in both precision and recall. Finally, we replace the threshold saliency maps in the CRF by the soft assignment approach of Section 3 (i.e. the second choice for f in (10)). Now, instead of sliding threshold τ we change the exponent κ, and achieve the corresponding precision-recall-curve in Figure 3 (black), which shows further improvement in performance.

In the best results were achieved by combining the band-pass saliency map with adaptive thresholding and the mean-shift segmentation algorithm. The achieved precision, recall, and F-measure values were 0.82, 0.75, and 0.78, respectively. The F-measure was computed from precision and recall by $F_\beta = (1 + \beta^2)(precision \cdot recall)/\$\beta^2 \cdot precision\ recall\%$, where $\beta = 0.3$ was used . This corresponds to a point marked using by a cyan star in Figure 3. This result remains lower than our results with both native thresholding and soft mapping with CRF, which provide the same precision with recalls 0.79 and 0.87, respectively. These points are also marked in Figure 3 with correspondingly colored stars. The maximum F-measure value we achieve is 0.85, which represents 9 percent improvement over The comparison of F-measures is shown in Figure 3. A few results of the proposed saliency segmentation method are shown in Figure 4 for subjective evaluation

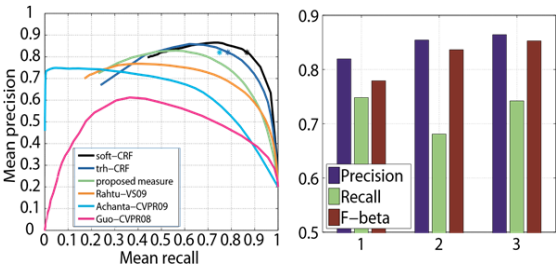


Fig3. Left: Mean precision-recall curves using comparison methods and the proposed approach. Right: Mean precision, recall, and F-measure values for comparison method (1), our method with thresholding (2), and our method with soft assignments (3). Notice that β = 0.3 (used according to) strongly emphasizes precision.

3.2 Saliency maps and Segmenting salient objects from images

Another set of experiments was performed using videos. The saliency maps were computed as described in Section 2 by using both the CIE Lab color values (only L in the case of gray-scale videos) and the magnitude of optical flow as features. which can provide real-time performance. The final salient segments were computed using either direct thresholding or the CRF method of Section. The results are compared with methods from which the last mentioned is a general background subtraction method. All comparison methods used default parameters given by the authors. Further, in order to achieve best possible performance with comparison methods, we also included all the post processing techniques presented in the original papers. As test videos, we used the publicly available image sequences. The two sequences from illustrate moving and stationary objects in the case of a fixed and a mobile camera. Sequences from show highly dynamic backgrounds with targets of various sizes. The original results of are available on-line and are directly comparable to our results. Their experiments also include several traditional background subtraction approaches.

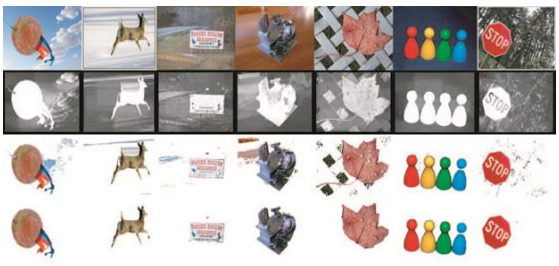


Fig.4 Examples of saliency maps and segmentations, Top row show the original image, second row shows the

saliency maps, third row shows the segmentations using threshold 0.7, and bottom row shows the segmentations using the CRF model.

IV. PROPOSED SYSTEM

In this paper, a new salient object detection framework for videos is proposed combining both spatial and temporal background priors. Specifically, since the background part is rigid for most natural videos, the temporal background priors are obtained based on the background homography between frames. The homography is estimated on point correspondences by means of SIFT flow in two frames. The spatial background priors are achieved utilizing the algorithm proposed in state-of-art, which are combined with abovementioned temporal background priors to generate the spatiotemporal background priors. For saliency value computation, a specially-designed motion-based graph is introduced to highlight uniform and accurate salient objects.

In summary, the contributions of this paper are as follows. Particularly for videos, besides image boundary, the temporal information (relation and difference among adjacent frames in chronological order) is also a significant hint, which needs to be considered especially for those videos that include complex scenes. Inspired by this idea, the temporal background priors are proposed as the complement of original definition "background prior"(in essence a more appropriate name should be "spatial background priors"). Combining these two concepts, we propose the "universal" spatiotemporal background priors to take full advantage of video information, by which an efficient and reliable salient object detection method for videos is built.

When producing final saliency maps, more and more graph-based models were proposed in recent years, in which graph construction is vitally important. However, they only take into account the spatial information of an image or video. This is not sufficient for describing the relationship among different nodes of the whole graph. In this work, we construct both an appearance graph and a novel motion graph, which simultaneously considers spatial and temporal relationship of different homogeneous objects.

The proposed framework is novel and is first introduced into the field of salient object detection for videos (video saliency).

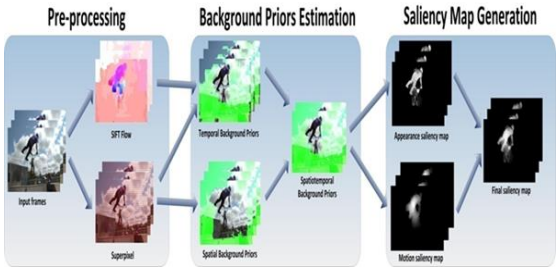


Fig 5: Framework of the proposed approach

4.1 Our approach in system architecture

Fig.5 shows the overview of the proposed salient object detection method with spatiotemporal background priors for videos. Initially, the pre-processing step is conducted to get superpixels as the basic unit of our approach and SIFT flow. The **scale-invariant feature transform** (**SIFT**) is a feature detection algorithm in computer vision to detect and describe local features in images. It was patented in Canada by the University of British Columbia and published by David Lowe in 1999.

4.2 Temporal background prior's estimation

Next, to estimate background priors, we first use the SIFT flow of all frames to obtain temporal background priors; spatial background priors are further achieved utilizing the algorithm proposed in. The above two types of background priors are combined to generate the spatiotemporal background priors. In saliency map generation, with these spatiotemporal background priors, we use the information of color and motion to compute appearance saliency maps and motion saliency maps respectively. These two types of saliency maps are then fused to produce final saliency maps.

SIFT key points of objects are first extracted from a set of reference images and stored in a database. An object is recognized in a new image by individually comparing each feature from the new image to this database and finding candidate matching features based on Euclidean distance of their feature vectors. From the full set of matches, subsets of key points that agree on the object and its location, scale, and orientation in the new image are identified to filter out good matches. The determination of consistent clusters is performed rapidly by using an efficient table implementation of the generalized Hough transform. Each cluster of 3 or more features that agree on an object and its pose is then subject to further detailed model verification and subsequently outliers are discarded. Finally the probability that a particular set of features indicates the presence of an object is computed, given the accuracy of fit and number of probable false matches. Object matches that pass all these tests can be identified as correct with high confidence

4.3 Initial temporal background priors estimation using SIFT flow and background homography

SIFT flow is a relatively new method to align an image to its nearest neighbors in a large image corpus containing complex scenes . It can match densely sampled, pixel-wise SIFT features between two images with the preservation of spatial discontinuities. Compared with traditional optical flow methods, SIFT flow does not need the assumptions of brightness constancy and piecewise smoothness of the pixel displacement field. It is robust to sudden illumination variation and scene changes. So we use SIFT flow to obtain the point correspondence between two frames. With the help of SIFT flow, the motion trend for each

point can be estimated. Mathematically, suppose we have a video V = fF ^{1}; F ^{2}; :::; F ^{N} g, where F ^{t} is the tth frame of the video and N is the total number of frames. For each pixel p$^{j}_{i}$ 2 F ^{j} where i is the pixel index, we can find its correspondence p$^{k}_{i}$ 2 F ^{k} by using SIFT flow between F ^{j} and F ^{k}.

Next, the homography for background will be estimated based on the assumption that the background part is rigid for most natural videos. Essentially, the background homography actually reflects the motion tendency of the majority of points in the scene from one frame to another adjacent frame. In detail, suppose the coordinate of p$^{j}_{i}$ is (x$^{j}_{i}$; y$_{i}^{j}$) and that of its correspondence p$^{k}_{i}$ is (x$^{k}_{i}$; y$_{i}^{k}$). The background homography H$_{j;k}$ is a 3 3 matrix for approximating the planar transformation between two adjacent frames F$_{j}$ and F$_{k}$, So if the aforementioned error err j;k(pji) is less than the threshold, it means the motion tendency of point pji agrees with that of the background and pji can thus be regarded as the initial temporal background prior. In the same way, if the error is bigger than the threshold, pji tends to have different movement trajectory from the background and is therefore more likely to be part of the foreground. This process can be formulated as

$$\underbrace{BP_{T(initial)}(p_i^j)}_{pixel\ level} = \begin{cases} 0, & if \left\| err_{j,k}(p_i^j) \right\|_2 \geq thres_1 \\ 1, & if \left\| err_{j,k}(p_i^j) \right\|_2 < thres_1 \end{cases}$$

Where BP$_{T\ (initial)}$(p$^{j}_{i}$) represents the initial status of the temporal background prior for the ith pixel in the jth frame. thres$_{1}$ is set to 1, which constrains the disparity between the estimated coordinate and the real coordinate within the distance of one-pixel unit so as to get more accurate initial temporal background priors.

4.4 Graph-based saliency detection

As is known, spatial and temporal information are two kinds of foremost information for video analysis. However, to the best of our knowledge, these two aspects haven't been used simultaneously in graph-based salient object detection. Traditionally, it is only the spatial information that is extracted to generate the graph for the computation of saliency values. In this section, driven by this point we propose a graph-based saliency method to make the best of both spatial and temporal information on superpixel level, which is especially suitable for salient object detection in videos. The whole process of saliency computation is shown in Fig. 3 using one frame instance.

4.4.1 Appearance saliency

In this part, appearance saliency is computed by constructing a graph, which measures the geodesic distance to background priors. In detail, firstly an undirected weighted graph G = fV; "g is created, where V represents the set of all nodes in this graph and each node corresponds to one superpixel. Besides, there exists an edge between every two adjacent nodes and the set of all edges are denoted by ". Each edge has a weight, which is defined as the color difference between two nodes and this difference is measured in Euclidean distance. It is worth mentioning that the color of each node is represented by the mean value of all pixels in its corresponding superpixel in R, G, B channels, respectively. The geodesic distance refers to the accumulated edge weights along the shortest path between two nodes in a graph model. It can measure the similarity between two nodes in the graph

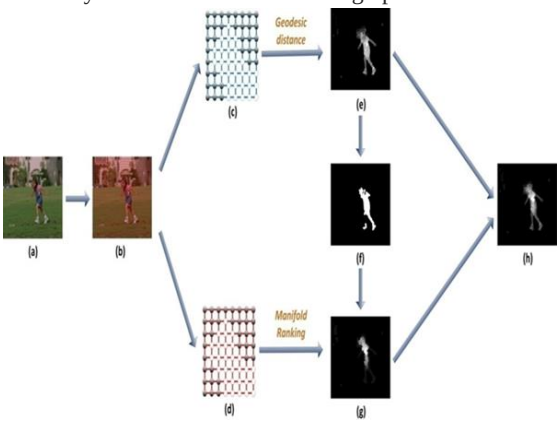


Fig 6: One frame instance to show the computation of saliency values based on the obtained spatiotemporal background priors from Section 3.3: (a) the input frame in a video; (b) extracting superpixels from (a); (c) and (d) are the graphs constructed by appearance and motion information respectively, in which black dots represent background prior super pixels; (e) is the appearance saliency map based on geodesic distances; after setting a threshold, the appearance-based binary salient region (ABSR) as query can be obtained like (f); by using manifold ranking, we can get motion saliency map (g) from (d) and (f); final saliency map (h) is generated as the fusion of (e) and (g).

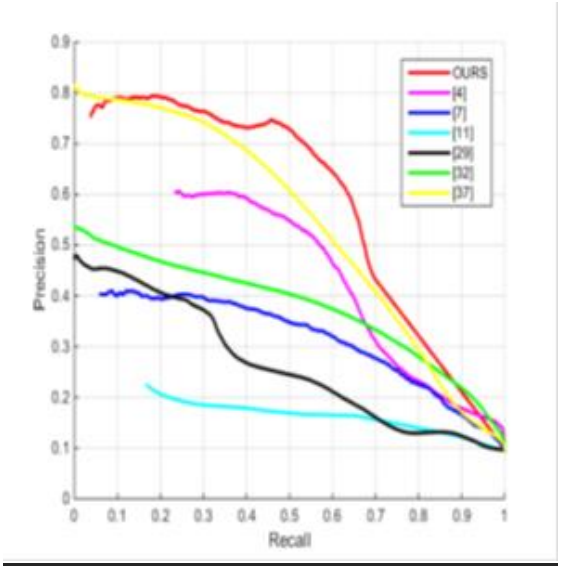
Model of an image with the above declaration, the appearance saliency for each superpixel in the jth frame is the geodesic distance between this superpixel and its nearest spatiotemporal background priors in the jth frame.

4.5 Experiments and Analysis

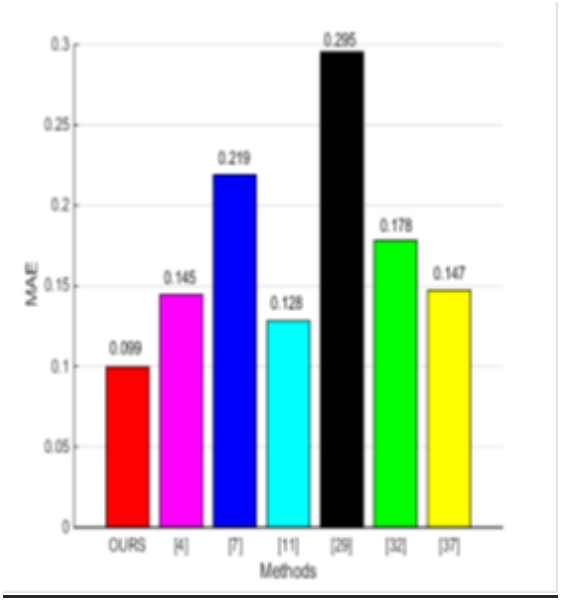
n this section, the proposed approaches are compared and evaluated for effectiveness and robustness against several state of- the-art saliency methods.

Among them, [7] is designed for image saliency detection; aim at video saliency detection. Using the source codes that corresponding authors provide, we obtain the saliency maps of above methods respectively. Also during experiments, different datasets are used, including both popular datasets with simple scenes and difficult datasets with complex scenes. A series of comprehensive comparisons and demonstration both

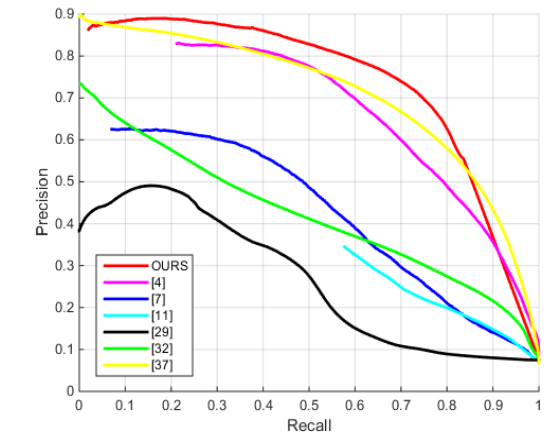
quantitatively and visually prove the validity and superiority of the proposed method.



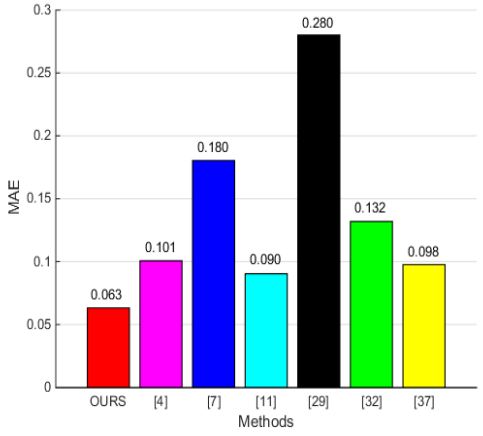
(a)



(b)



(c)



(d)

Fig 7: Quantitative comparisons between the proposed method and other six existing approaches using SegTrack and SegTrack v2 datasets as the benchmark of simple scenes: (a) and (c) are precision-recall curves for SegTrack and SegTrack v2 respectively by setting the thresholds from 0 to 255 for obtained saliency maps; (b) and (d) are mean absolute errors (MAE) for SegTrack and SegTrack v2 respectively.

4.6 Datasets with simple scenes: SegTrack and SegTrack v2

SegTrack [35] and SegTrack v2 [26] are the most popular benchmark with relatively simple scenes in the fields of video saliency, tracking, segmentation and so on. SegTrack is a video segmentation dataset with full pixel-level annotations on one or multiple objects at each frame within each video. It contains six videos with pixel-wise ground truth, namely “Birdfall”, “Cheetah”, “Girl”, “Monkey dog”, “Parachute” and “Penguin”. “Birdfall” was shot by stationary camera and other remaining videos were taken using moving cameras.

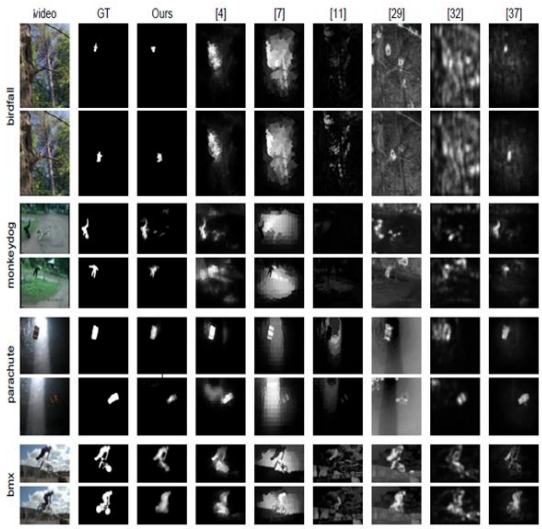


Fig. 8: Visualization effects for the compared methods using SegTrack and SegTrack v2 as the benchmark of simple scenes.

Besides, each frame of a video in SegTrack corresponds to one binary mask as the ground truth. SegTrack v2 can be seen as an updated version of SegTrack. Different from SegTrack ,SegTrack v2 consists of 14 videos under more challenging circumstances, like appearance change, motion blur, occlusion, complex deformation, interacting objects, slow motion and so on. Note that in SegTrack, the ground truth for multiple salient objects is incomplete and incorrect: some only contain one object as the ground truth, like "Monkeydog", "Penguin" and "Cheetah". So for the above videos in SegTrack, we use the ground truth from SegTrack v2 for evaluation, which is complete and correct. It is worth mentioning that many researchers skip "Penguin" video and only use the remaining five videos for test because of the incomplete ground truth and also complexity. Here we use all videos to keep the objectivity and impartiality of experimental results for a complete dataset. To evaluate the all-round performances and disparities between our method and other existing ones, we test the experimental results based on both quantitative criteria and visualization effects. For the former, two popular metrics are used, including precision-recall (PR) curve and mean absolute error (MAE). Specifically in PR curve, "precision" refers to the percentage of salient pixels which are allocated correctly in the obtained saliency maps; "recall" represents the percentage of detected salient pixels. For plotting the curve, obtained saliency maps are binarized with a series of thresholds from 0 to 255. Then 256 pairs of precision-recall combinations are generated and the curve can be drawn. From the implication of PR curve's definition, the ideal condition is that both precision and recall are equal to 1, so the curve closer to (1; 1) (top right corner) corresponds to better performance. Another assessment criterion is the mean absolute error (MAE) [36], which aims at a more balanced comparison between the binary ground truth GT and the continuous saliency map S for all frame pixels. It can be defined as

$$MAE = \frac{1}{N_p} \sum_{i=1}^{N_p} |GT(i) - S(i)|$$

where Np is the number of frame pixels and i is the pixel index. As can be seen from the above expression, MAE depicts the degree of approximation between the obtained continuous saliency map (normalized to the scale of [0; 1]) and binary ground truth. Smaller MAE means smaller dissimilarity and better performance. Fig. 4 shows the resulting quantitative comparisons between the proposed methods and other six approaches when Seg-Track and SegTrack v2 are used to evaluate the performance of simple scenes. As can be seen in Fig. 4(a)(c), the PR curve describes the degree to which saliency maps highlight salient objects (regions). Higher PR curve which is closer to (1; 1) corresponds to better performance. For the most range, our method ranks the highest, which shows that the proposed algorithm outweighs other listed methods to a great extent. It is noteworthy that the proposed method achieves the best precision rate up to 0.83 in SegTrack. For SegTrack v2, although more videos with more challenging circumstances are included, this number still remains a high (or even higher) level which reaches up to 0.9. Besides, when recall=0.6, precision still remains at a high level, which is above 0.7 and 0.8 respectively in SegTrack and SegTrack v2. This indicates our saliency maps are more precise than other existing methods while maintaining a high-level response to salient regions. More intuitionally from Fig. 4(b)(d) in terms of MAE Our method improves the performance by 22% and 30% respectively in SegTrack and SegTrack v2 when compared to the smallest value of other compared methods. This indicates the validity and superiority of the proposed method in the case of simple scenes. For further analysis, other existing methods mostly use the information from image/frame boundary. It may result in inaccuracy when salient objects are located near the boundary region. Consequently,

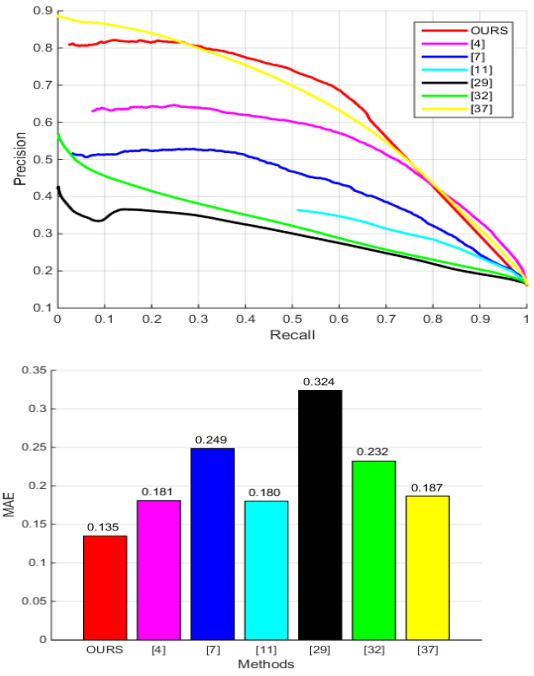


Fig.9: Quantitative comparisons of PR curve and MAE between the proposed method and other six existing approaches using FBMS-59 datasets as the benchmark of complex scenes.

these methods are likely to lose some portions of salient objects. In contrast, our success stems from that 1) we use effective and accurate background priors to distinguish the salient object from the video; 2) our saliency maps are more uniform. This point is attributed to the utilization of motion saliency. It makes the portions of similar motions share corresponding similar

saliency values, which lets the salient object more uniform. Fig. 5 provides the comparisons of visualization effects for different methods. As can be clearly seen, the proposed method achieves optimal performance, which is nearest to the ground truth when compared with other six methods. The defects of these methods can be summarized into several categories: 1) the detected result is incomplete and only parts of the salient object are detected; 2) parts of the background are wrongly detected as components of the salient objects; 3) the detected salient object contains too much noise; 4) salient objects cannot be detected. However, in contrast, our results show two important characteristics of detected salient objects: 1) Accuracy. In comparison to other traditional approaches where the image boundary is used as background priors (such as [31]), we take full advantage of information from both spatial and temporal domain to obtain the better and more robust background priors. 2) Integrity. We use the information of motion saliency, which makes all parts of the salient object be detected as a homography.

Datasets with complex scenes: Freiburg-Berkeley Motion Segmentation Dataset

Freiburg-Berkeley Motion Segmentation Dataset (FBMS-59) is a popular and more challenging dataset in the field of moving object segmentation and video saliency. It is a large benchmark with 59 heterogeneous video sequences, many of which contain complex scenes. So it is an appropriate dataset for evaluating the performance of different salient object detection methods in tough and complicated situations. Note that FBMS-59 is a rather big dataset including both training sets (29 videos) and test sets (30 videos); besides, the ground truth images are provided only for a small proportion of frames instead of all frames. So for efficiency and handy processing, we select its test set with the corresponding given ground truth for experiment. Same methods of comparison and assessment criteria are utilized as those in Section 3.7.1. The experimental results are shown in Fig. 3.7.1. For Fig. 3.7(a), in most ranges, the PR curve of the proposed method is the highest and closest to (1; 1); for Fig. 6(b), our MAE is the lowest and obtain a 25% improvement against the second lowest method. These two quantitative comparisons show that even when video scenes become more complex, the proposed method can still achieve the best performance in all compared approaches. These results further prove that our method is accurate and robust against different scenarios. Fig.3.7.2 demonstrates some typical kinds of complex scenes. As can be seen from Fig 3.7.2(a), both the white bus and the red car are salient objects and our method can simultaneously detect these two; however, for remaining compared methods, the best result still fails to detect thoroughly. (b) is from a video in which a dog is running past a horse. This frame is just shot when partial body of the dog is overlapped by the horse's leg and it causes partial occlusion. Our method can precisely detect the dog while others can't.

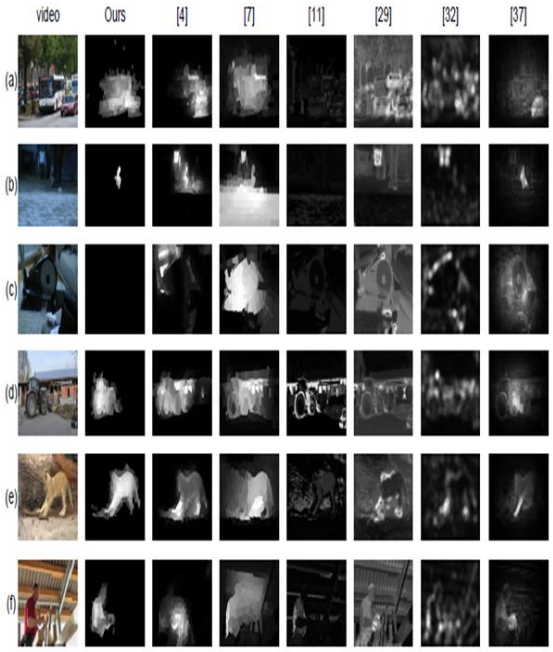


Fig.10: Some typical examples of complex scenes and the corresponding visualization effects using the compared methods: (a) multiple salient objects; (b) partial occlusion; (c) complete occlusion; (d) complex background; (e) color similarity between salient object and background; (f) light disturbance.

(c) is cut from a video where a rabbit is circuiting a elliptical machine and at this time the rabbit's whole body is entirely occluded by the machine. The visualization result of our method honestly reflect the truth: it cannot detect the salient object; meanwhile, others detect "false" salient regions. For (d), the background information is very complex: tangle some weeds, shabby factory with broken windows and doors. Under such tough circumstances, our
method can still detect all regions of the whole tractor and even the loader in the tractor is clear.

However, other compared ones detect more than that due to the interference of the complex background: a portion of the factory is wrongly deemed as the salient object. In (e), the salient object (lion) and the background share similar color in some regions, which is difficult to distinguish the foreground from the background and causes the failure of integrated detection. Nevertheless, since the proposed method utilizes motion saliency, it can make up for the deficiency when only appearance information is used. Thus, the proposed method can obtain an exact and intact detection. As for (f), an old man is repairing something on the ladder. Because the light filters through the slits around the man, other methods detect these illuminated regions as more salient and neglect the "authentic" salient object. By contrast,

our method can correctly represent the salient object with the existence of light disturbance. To sum up, all the aforementioned extreme cases in a more complex scene illustrate that the proposed method is capable of resisting various disturbance and maintaining the accurate and robust performance in a video containing complex scenes.

V.CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

In this paper, we have proposed a novel graph-based salient object detection approach for videos using spatiotemporal background priors. Firstly the temporal background priors are obtained based on the background homograph between frames, which are estimated by applying RANSAC on point correspondences utilizing SIFT flows of multiple-pair frames. Then, the spatial background priors are combined with above mentioned temporal background priors to generate the spatiotemporal background priors. Finally, by respectively constructing the appearance graph and motion graph, the saliency map for each frame is obtained after measuring the difference from the spatiotemporal background priors. Compared with other state-of-the-art methods, the proposed method achieves highest performance on different challenging datasets of both simple and complex scenes, which exhibits higher robustness and accuracy.

5.2 FUTURE SCOPE

In the project for future scope essential things are added, we done as best as possible but we concentrating on low quality videos and images to gain maximum accurate output of main object or required object of image or video and here we have been using good resolution cameras but we also concentrating in the software techniques to get the consistent output of the main object and we also thinking for future scope it will get more reliable and comfortable techniques to use SIFT frame work. Firstly the temporal background priors are obtained based on the background homograph between frames this give main object information but we take other best graphical mapping method for video effects in the future. That is here we have working in changing movements in 3d graphic video effects to get matching movements of objects. Applications adding for feature are include object recognition, robotic mapping and navigation, image stitching, 3D modeling, gesture recognition, video tracking, individual identification of wildlife and match moving and where ever video scanning require. Using different essential other updates to this practical technique we get accurate salient detection of objects in images as well as in videos for the above applications in feature.

REFERENCES

[1] Y. Qin, H. Lu, Y. Xu, and H. Wang, “Saliency detection via cellular automata”, in IEEE Conf. on Computer Vision and Pattern Recognition ,pp. 110-119, 2015.

[2] Z. Liu, W. Zou, L. Li, L. Shen, and O. L. Meur, “Co-saliency detection based on hierarchical segmentation”, IEEE Signal Processing Letters,21(1):88-92, 2014.

[3] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk,“SLIC super pixels compared to state-of-the-art superpixel methods”,IEEE Trans. on Pattern Analysis and Machine Intelligence, 34(11):2274-2282, 2012.

[4] W. Wang, J. Shen, and F. Porikli, “Saliency-aware geodesic videoobject segmentation”, in IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3395-3402, 2015.

[5] Y. Fang, W. Lin, Z. Chen, C.-M.Tsai, and C.-W. Lin, “A video saliency detection model in compressed domain”, IEEE Trans. on Circuits and Systems for Video Technology, 24(1):27-38, 2014.

[6] D. Zhang, J. Han, C. Li, and J. Wang, “Co-saliency detection vialooking deep and wide”, in IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2994-3002, 2015.

[7] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, “Saliency detection via graph-based manifold ranking”, in IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3166-3173, 2013.

[8] W. Zhu, S. Liang, Y. Wei, and J. Sun, “Saliency optimization from robustbackground detection”, in IEEE Conf. on Computer Vision and PatternRecognition, pp. 2814-2821, 2014.

[9] Y. Luo and J. Yuan, “Salient object detection in videos by optimal spatiotemporalpath discovery”, in ACM Intl. Conf. on Multimedia, pp. 509-512,2013.

[10] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum,“Learning to detect a salient object”, IEEE Trans. on Pattern Analysisand Machine Intelligence, 33(2):353-367, 2011.

[11] H. Fu, X. Cao, and Z. Tu,“Cluster-based co-saliency detection”, IEEETrans. on Image Processing, 22(10): 3766-3778, 2013.

[12] W. Kim, C. Jung, and C. Kim, “Spatiotemporal saliency detection andits applications in static and dynamic scenes”, IEEE Trans. on Circuitsand Systems for Video Technology, 21(4):446-456, 2011.