IJETIE
International Journal of Emerging Technology and Innovative Engineering
Volume I, Issue 3, March 2015
ISSN: 2394 - 6598
www.ijetie.org

SALIENCY APPROACH TO EXTRACT OBJECT FROM VIDEO via KINECT SENSOR

Kedar A. Bhartiya
Sir Visvesvaraya Institute of technology, Nashik
bhartiyakedar@gmail.com

Siddharth Pathak
Sir Visvesvaraya Institute of technology, Nashik
siddharth.pathak11@gmail.com

Monil P. Shah
Sir Visvesvaraya Institute of technology, Nashik
smonil335@gmail.com

Nitesh M. Nagdeote
Sir Visvesvaraya Institute of technology, Nashik
niteshn151@gmail.com

Prof. S.M. Rokade
HOD & Associate Proffessor
Sir Visvesvaraya Institute of technology, Nashik
smrokade@yahoo.com

ABSTRACT

The paper performs video object extraction (VOE) framework using the saliency approach. The framework aims to extract foreground objects of willing without any user interaction or the use of any training data automatically. To distinct foreground and background regions from the video frames, the preferred method exploits visual and motion saliency information extracted from the input video. The term conditional random field is solicited to combine the saliency induced features in effective manner apparently it permits to deal with unknown stance and scale variations of the foreground object. Our Proposed work does not require any prior knowledge on the object of interest or any interaction from the user. As another input medium we used the sensors of kinect device. This can realize the gesture based on the information of depth image get by kinect sensor. Experiments reveal the advantages of using the gesture recognition by kinect's depth sensor.

General Terms

Saliency-

It is also spells as Visual Salience. The distance subjective perceptual quality which marks some item in the world standout for their neighbors and immediately grab our attention. Our attention is attracted to visually salient stimuli. It is important for complex biological system to rapidly detect potential prey, predators, or mates in a cluttered visual world. However simultaneously identifying any and all interesting target in one's visual field has prohibitive computational complexity making it a daunting task even for the most sophisticated biological brains, let alone for any existing computer.

Visual Saliency-

It is also spells as Visual Salience. The distance subjective perceptual quality which marks some item in the world standout for their neighbors and immediately grab our attention. Our attention is attracted to visually salient stimuli. It is important for complex biological system to rapidly detect potential prey, predators, or mates in a cluttered visual world. However simultaneously identifying any and all interesting target in one's visual field has prohibitive computational complexity making it a daunting task even for the most sophisticated biological brains, let alone for any existing computer.

Motion Saliency

In order to exploit the temporal characteristics of the input video, we capture the motion information by calculating optical flow. To perform more accurate optical flow estimation, we apply dense optical-flow with both forward and background propagation at every frame. The decomposition of an object shape model in a hierarchical way to train object part detectors, and these detectors are used to describe all possible configurations of the object of interest (e.g. pedestrians). Another type of supervised methods requires user interaction for annotating candidate foreground regions.

Keywords

video object extraction, motion saliency, visual saliency, conditional random field, kinect device.

1. INTRODUCTION

The human being can certainly determine the willing subject from the video, even the subject is in very pseudo or faux condition or having cluttered background or even has never seen before. Due to complex cognitive capabilities exhibited by human brain it explicate as concurrently extraction of foreground and background information from the video frame. Lots of researchers are trying to proximate the gap between human brain capabilities and the computer vision. But, without any preliminary knowledge on the subject of interest or training data it is very tedious for computer vision to extract the foreground of object of willing automatically from the video. If anyone wants to design an algorithm to automatically extract the foreground objects from a video, several tasks need to be addressed.

1) Unknown object category and unknown number of the object instances in a video.

2) Complex or unexpected motion of foreground objects due to articulated parts or arbitrary poses.

3) Ambiguous appearance between foreground and background regions due to similar color, low contrast, insufficient lighting, etc. conditions.

In practical area, it is inconceivable to manipulate all foreground or background models beforehand. However, if one can uproot the constitute information from both foreground and background region from the video frame, the uprooted information can be exploit to distinguish between foreground and background region, and hence the procedure for extracting the foreground object can be figured out. Many of the prior works either consider a fixed background or assume that the background exhibits dominant motion across video frames [1]. These assumptions might not be practical for real world applications, since they cannot generalize well to videos captured by freely moving cameras with arbitrary movements.

This paper proposed a robust framework for the video object extraction (VOE), which utilizes visual as well as motion saliency information over the video frames. The observed saliency information allows us to infer several visual and motion cues for learning foreground and background models, and a conditional random field (CRF) is applied to automatically determines the label (foreground or background) of each pixel based on the observed models. With the great potential to preserve the spatial and temporal consistency, our framework for the VOE exhibits assurance of result on a variety of videos, and generate the quantitatively and qualitatively adequate performance [2]. While we concentrate on the problems occurring in VOE for the single concept videos, our preferred method is capable to deal with multiple object instances (of the same type) with pose, scale, etc. variations.

This paper proposed an another input medium called kinect to the system for the further working process, which can perceive the gesture track recognition based on the depth image information get by the kinect sensor [3]. At first, the kinect sensor is used to obtain depth image information, and then it extracts splith. Many experiments exhibits the advantages of using the gesture recognition of the kinect's depth image can be very effective to achieve interactive features.

Traditionally the gesture recognition based on color image has gained substantial achievements, but, it can't describe the depth information of the gesture since it lacks the three dimension parameter. However, the depths chart to make up for this shortcoming has been widely applied for the study of the gesture trajectory tracking, basically kinect is a part of the Xbox 360 and windows video game control platform developed and designed by Microsoft. The range of Infrared combined with an RGB camera on the device detects the gesture motion of human.

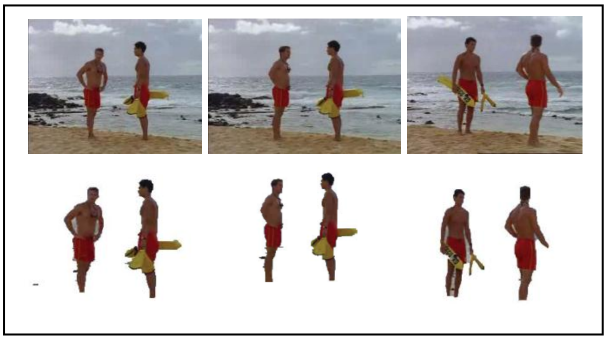

Fig. 1: Overview of our proposed framework for VOE.

2. EXISTING SYSTEM

For the separation of foreground and background regions over the video frames, it exploit the information of visual and motion saliency extracted from the input video, Lin and Davis [4] both decomposed an object shape model in a hierarchical way to teach object part detector, and these trained detectors are used to represents all possible configurations of the object of willing. Another type of extraction methods requires user interaction for annotating candidate foreground regions. For example, image segmentation algorithms proposed in [5], [6] focused on an interactive scheme and required users to manually provide the ground truth label information. For videos captured by a monocular camera, methods such as Criminisi *et al*., Yin *et al*. [7], applied(CRF) maximizing a joint probability of color, motion, etc. The model is independent of features, categories, or other forms of prior knowledge of the objects. An interactive framework for soft segmentation and matting of natural images and videos is presented. The focus of automatic detection of visually salient regions in images that describes a new method for automatic target segmentation [8] and tracking which uses a multi-label Markov Random Field (MRF) formulation to sequentially "carve" a target of interest out of a video volume.

3. RELATED WORK

Formally there are two approaches used to figure out the problem of VOE i.e. supervised and unsupervised approach. Supervised methods require prior knowledge on the subject of interest and need to collect training data beforehand for designing the associated VOE algorithms. As per Lin and Davis [4] both decomposed an object shape model in a hierarchical way to train object part detectors, and the detectors decomposed by them are used to describe all possible configuration of the object of interest. The supervised method can also be performed in other ways; it required the user attention for annotating candidate foreground regions. In segmentation algorithm it focused on an interactive scheme and required users to manually provide the ground truth label information. The videos taken by monocular camera, the Criminisi *et al*., Yin *et al*. [5], [6] methods are appealed a conditional random field (CRF) which maximizing a joint probability of color, motion, etc. models to predict the label of each image pixel. While working on this method the color features can be automatically examine from the input video, but this methods still need the user to train object detectors for extracting the other features like shape and motion. Resent researches presented some preliminary strokes which helps manually selection of the foreground and background regions, and they exploited the information to the local suspects to detect the foreground objects. But it is a tedious job and it might not practical for user to annotate a huge amount of data manually from video.

In case of the unsupervised approach it does not train any particular object detectors or classifiers. The extraction of foreground objects for the videos taken by the static camera treated as a background subtraction problem. It also can be stated as; the objects in foreground can be detected simply by subtracting the ongoing video frame. But in many situations the background is consistently changing or may be covered by the foreground objects, in that case background modeling becomes a very tedious and challenging task. For such type of cases, researchers particularly aim at learning the background model from the input video, and the objects lied in foreground are considered as outlier that to be detected. For example Sun *et al*. [9] utilized color gradients of the background to determine the boundaries of the foreground objects. Some unsupervised approaches aim at observing features associated with the foreground object for VOE.

In concern of the kinect device many scholars are fascinated to use the kinect, and the research about the acquisition of depth image using the gesture also gives the substantial results. Zhang [10] tested the kinect with a hidden Markov model (HMM) in way of achieving gesture trajectory recognition and Wu [11] tested skeleton point cloud. Van Bang Le *et al*. [3] proposed the study about the use of image processing library of pykinect SDK and of OpenCV to process of Kinect depth image and get accurate hand segmentation.

Depending on the relationship between the gray level of depth image and the actual distance, it is easy to detect the hand contour portion, and after that calculate the center of gravity to find out the position of the hand. It is quite simple method of segmentation than the traditional visual inspection methods.

4. PROPOSED SYSTEM

In the proposed work the aim is to automatically extract the foreground object from videos which are taken by the freely working cameras and the kinect device. Rather assuming that background motion is presiding and vary from the foreground, in proposed work this assumption barely relax and allow the foreground object to be presented in freely working scenes. The suggested work enriches the visual and motion saliency information over the video frames, and the conditional random field model is exploit for integration of the associated features for VOE. From the quantitative and qualitative experiments, It is verified that the proposed VOE performance exhibits spatial consistency and temporal continuity, and its presents the outperform state-of-art unsupervised VOE approaches. It is beneficial that the proposed VOE framework in an unsupervised approach, that does not require any prior knowledge of the suspect of willing also not any need of user interaction anymore. Also the segmentation method employs both visual and motion cues, and it combines dynamic information and spatial interaction of the observed data[11]. Experimental results show that the proposed approach effectively fuses contextual constraints in video sequences and improves the accuracy of object segmentation.

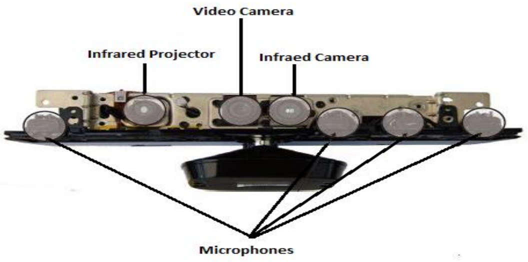

Fig. 2: A Kinect sensor unwrapped.

In addition as our contribution we present to get the video frames from the Kinect device. Fig. 2 shows a Kinect with the internal structure of it that included a infrared projector, infrared camera, RGB camera and microphones[3]. Infrared camera to detect environment space depth output for 640×480 resolution video (low frame rate can be up to 1280×1024 resolution). From the time it is used with Xbox 360 gaming platform is connected, the available range of 1.2 ~ 3.5 meters[3][4].

The collected data from the kinect sensor is exhibits in Fig.3. In depth image, the distance between the human or object body and the IR camera is proportional to the grayscale. Infrared rays on an object's surface reflection to the receiver to form a three-dimensional coordinates of an object point cloud. So every frame from the video has slightly different contour information, and not able to solve the problem of shade and it also, for some special materials or special surface structure, will produce a lot of noise[4].

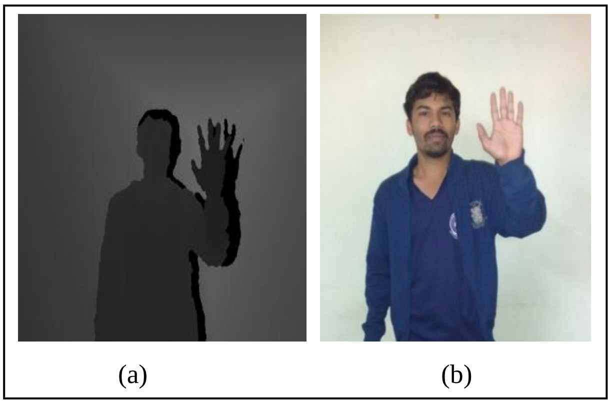

Fig. 3: Sampling data of Kinect. (a) Depth image, (b) RGB image.

5. MODULE DESCRIPTION

5.1 Automatic object extraction

In many of the existing unsupervised VOE approaches it assumes the object at foreground as outlier in terms of the observed motion information, so that the induced appearance, color etc. features are exploited for differentiate between the foreground and the background regions. Yet, as we consider early, these methods cannot generalize well to the videos which are taken by the freely working cameras[4][5]. In the work of object extraction we exhibit a saliency-based framework which grabs the saliency information in both visual and motion domains. By approaching conditional random field (CRF), the unification of the resulting features can automatically recognize the object at foreground without need to treat either foreground or background as outliers. Fig. 4 shows the referred VOE framework, and we now detail each step as following.

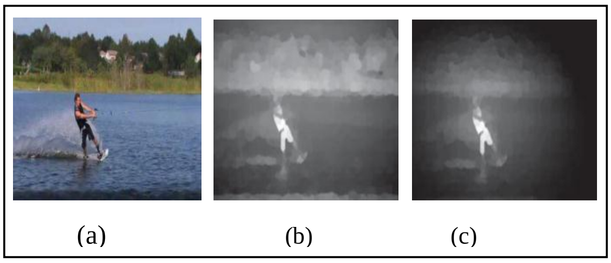

Fig. 4: Example of visual saliency calculation. (a) Original Video frame. (b) Visual saliency of (a) derived by (1). (c) Refined by (b)

5.1.1 Determining the visual saliency

To extract visual saliency from each frame, we perform image segmentation on each video frame and extract color and contrast information. While working we advanced the Turbopixel for the segmentation [8], and the resulting superpixel (segments) are enforce in detection of the saliency. The Turbopixel helps to produce edge preserving superpixels with homogeneous sizes, which leads to achieve the improved visual saliency results as verified further. For the *k*th superpixel *rk*, we calculate its saliency score *S(rk)* as follows:

$$S(rk) = \sum_{rk \neq ri} \exp(Ds\,(rk, ri)\sigma_s^2 \omega(ri) Dr\,(rk, ri)$$

$$\approx \sum_{rk \_= ri} \exp(Ds\,(rk, ri)/\sigma_s^2) Dr\,(rk, ri)\ldots..(1)$$

The last term *Dr (rk, ri)* measures the color difference between *rk* and *ri*, which is also in terms of Euclidean distance. Then we consider the pixel *i* as a silent point if that point score satisfies *S(i)* >0.8 * max(*S*), and the bunch of concluding salient pixels will be considered as a salient point set. Since image pixels which are closer to this salient point set should be visually more significant than those which are farther away, we further refine the saliency,

$$\widehat{S}(i) = S(i) * (1 - \text{dist}(i)/\text{distmax}) \ldots\ldots(2)$$

Where *S(i)* is the original saliency score derived by (1), and dist*(i)* measures the nearest Euclidian distance to the salient point set[3][5]. We derived that distmax in (2) is determined as the maximum distance from a pixel of willing to its nearest salient point within an image, and hence it is an image-dependent constant. An example of visual saliency calculation is shown in Fig. 4.

5.1.2 Determination of motion saliency

Now lets us see that how to analyze or determine the motion saliency, and how we extract the associated cues for VOE purposes. Unlike previous work which assume that either foreground or background shows the presiding motion, in our referred framework we aims toward the extraction of motion salient regions based on the retraced optical flow information. For detecting each part and its corresponding pixel, we conduct the murky optical-flow forward and backward propagation at each frame of the video input. A moving pixel *qt* at frame *t* is determined by,

$$qt = \widehat{qt}, \mathbf{t} - \mathbf{1} \cap \hat{\mathbf{q}}t, t+1 \ldots\ldots\ldots\ldots\ldots(3)$$

Metric \ Methods	CA [22]	LD [20]	ST [23]	MSSS [21]	HC [27]	RC [27]	Ours
F-measure	0.9100%	0.8667%	0.6222%	0.7839%	0.7032%	0.8067%	0.8617%

Where $\hat{q}$ denotes the pixel pair detected by forward or backward optical flow propagation[1][2]. We are not able to avoid the frames which result in a huge number of moving pixels, and thus our setting should be more practical for the purpose of real-world videos which are taken by the freely working cameras. After calculating the moving region, for the each pixel in the terms of the associated optical flow information we refer to derive the saliency score. As likely as the visual saliency approaches we perform our refer algorithm in equations (1) and (2) on the derived optical flow which calculate the results in motion saliency M*(i,t)* for the every pixel *i* at the frame *t*, and the saliency score at every frame is normalized to the range of[0,1]. It is stated as, when the object at the foreground shows the significant movements as compared to the background ,its motion will easily grabbed by the optical flow and hence the corresponding motion salient region get extracted easily. On the other side if the camera is in moving condition and thus result in remarkable background movements, the motion saliency method which is preferred will still able to identify the motion salient regions.

5.1.3 Conditional random field

While defining the conditional random field lets us go through the assumption, For two random fields r and r′ over the video scene, (r, r')is a conditional random field if, when conditioned on r′, the random field r obey the Markov property (r(x)|r', r(y), y≠ x) = P (r(x) | r', r(y), y ∈ N_x) denotes the neighboring sites of point x. Using the Hammersley-Clifford theorem and considering only up to pair wise clique potentials, the posterior probability of the segmentation field is calculated.

To calculate the temporal dependencies of the adjacent segmentation fields, the state transition probability of the segmentation field p(sk+1 | sk) is modeled using a Gibbs distribution defined on one-pixel and two-pixel cliques as well. The observation (or likelihood) model p (zk | sk) is also formulated by a conditional random field to capture dependencies between observations. Using the spatial and temporal [8] dependencies in the process of segmentation are combined by a dynamic probabilistic framework based on the CRF model. On the basis of the dynamic framework formation the new concept is formed i.e. Dynamic conditional random field DCRF). The DCRF extends the CRF for individual images by incorporating temporal dependencies among segmentation field.

Table 1. Comparisons of maximum f-measure scores for different visual saliency detection approaches

5.2 Augmented Reality

Basically the Augmented reality is a system which has a goal to enhance the user's perception of and the interaction with the real world through supplementing the real world with 3D virtual objects which appear to co-exist in the same space as the real world. Many researchers have worked on the topic of AR beyond the new level of the vision, but in spirit of the original survey we define the AR system to share the following properties:

1) Blends real and virtual, in a real environment
2) Real-time interactive
3) Registered in 3D

The Registration mentioned to the accurate alignment of the real and virtual objects[12]. Without the precise registration, the illusion that the object which is in virtual form exist in the real environment is severely compromised. There are many different definitions of AR so it is not restricted to particular display technologies, like as neither Head-Mounted Display (HMD) nor it has the limitations with the visual sense[11]. AR can potentially apply to all senses, like as touch, hearing etc. Many of the AR applications also needed to remove real object from the environment, in addition to adding virtual objects. Augmented reality allows gamers to experience digital game play in a real world environment. In the last 10 years there have been a lot of improvements of technology, resulting in better movement detection and the possibility for exist, but also direct detection of the player's movements.

Many of the AR interfaces were based on the desktop metaphor or used design from virtual environments research. Among many of the trends one main trend in interaction research notably for AR system is the use of Heterogeneous designs and tangible interface. Heterogeneous approaches blur the boundaries between real and virtual, taking parts from both worlds. Tangible interfaces emphasize the use of real, physical objects and tools. Since in AR systems the user sees the real world and often desires to interact with real objects, it is appropriate for the AR interface to have a real component instead of remaining entirely virtual.

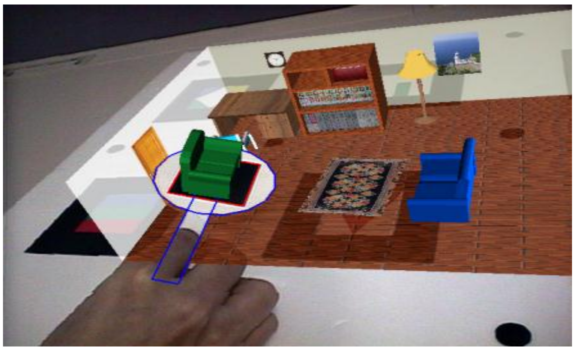

Fig. 5: User wields real paddle to pick up, move, drop, and destroy models.

The above example is of such an interface, the user wields a real paddle to handle furniture models in a prototype interior design application Through pushing, tilting, swatting and other motions, the user can select pieces of furniture, drop them into a room, push them to the desired locations, and smash them out of existenceto eliminate them (Fig. 5).

6. ACKNOWLEDGMENTS

Perseverance, Inspiration and motivation have always played a key role in the success of any venture. At this level of understanding, it is often difficult to understand a wide spectrum of knowledge without proper guidance to our respected project guide. It is indeed, a great pleasure and a moment of immense satisfaction for us to express our profound towards our guide.

We extend our thanks to Prof.S.M.Rokade, Our guide and Head of Computer Engineering Department, for providing us the opportunities & facilities and inspiration, gather professional knowledge and material to complete this project, without which it would have been impossible to complete this hard task.

7. REFERENCES

[1]. Wei-Te Li, Haw-Shiuan Chang, Kuo-Chin Lien, “Exploring Visual and Motion Saliency for Automatic Video Object Extraction”, ieee transactions on image processing VOL. 22, NO. 7, JULY 2013.

[2]. Nithin.P.Varkey, Prof.S.Arumugam, “Automatic video object extraction by using generalized visual and motion saliency”, international journal of innovations in scientific and engineering research (ijiser), p-issn: 2347-9728, vol 1 issue 4 apr 2014.

[3]. Van Bang Le, Anh Tu Nguyen, and Yu Zhu, “Hand Detecting and Positioning Based on Depth Image of Kinect Sensor”, International Journal of Information and Electronics Engineering, Vol. 4, No. 3, May 2014.

[4]. Z. Lin and L. S. Davis, “Shape-based human detection and segmentation via hierarchical part-template matching, IEEE Trans. Pattern Anal.Mach. Intell, vol. 32, no. 4, pp. 604–618, Apr. 2010.

[5]. J. Sun, W. Zhang, X. Tang, and H.-Y. Shum, “Background cut,” in Proc. 9th Eur. Conf. Comput. Vis., 2006, pp. 628–641.

[6]. Y. Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[7]. A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov, “Bilayer segmentation of live video,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2006, pp. 53–60.

[8]. Wei-te li, hui-tang chang, hermes shing lyu, and yu-chiang frank wang, “Automatic saliency inspired foreground object extraction From videos”, icip 2012.

[9]. J. Sun, W. Zhang, X. Tang, and H.-Y. Shum, “Background cut,” in Proc. 9th Eur. Conf. Comput. Vis., 2006, pp. 628–641.

[10]. Y. Zhang, S. Zhang, Y. Luo, and X. D. Xu, “Gesture trajectory identification and application based Kinect depth image,” Application research of computer, 2012, vol. 29, no. 9.

[11]. S. Wu, F. Jiang, and D. B. Zhao, “Hand gesture recognition based on skeleton of point clouds”, 2012 IEEE fifth International Conference on Advanced Computational Intelligence (ICACI), pp. 566-569, October 2012.

[12]. Ronald Azuma, Yohan Baillot, Reinhold Behringer, Steven Feiner, “Recent Advances in Augmented Reality”, Computers & Graphics, November 2001.