ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

Saliency Based Video Object Recognition in Single Concept Video

Raihanath A.S [1], Chithra Rani P R [2]

[1] P G Student, Department of Computer Science, Ilahia College of Engineering & Technology (ICET), Kerala, India

[2] Assistant Professor, Department of IT, Ilahia College of Engineering & Technology (ICET)), Kerala, India

ABSTRACT: This paper presents a saliency based video object extraction and recognition framework. The extraction framework automatically extract foreground object of interest without any use of training data. The recognition system uses list of training data beforehand. For extracting the foreground object from a video, it uses visual and motion saliency features .A conditional random field is used to effectively combine the saliency induces features. Our proposed method is able to preserve spatial continuity and temporal consistency. Experiments results on variety video shows that our proposed system provides qualitatively satisfactory video object extraction (VOE) results.

KEYWORDS: Video Object Recognition, Visual Saliency, Conditional Random Field.

I. INTRODUCTION

Human can easily interpret the salient object from a video, using the capabilities of human brains. But, in computer vision it is very challenging to recognise the salient object. Researchers are trying to close the gap between the computer and human vision. It is very challenging for a computer vision algorithm to automatically extract the foreground object from a video without any of human interaction. However, if one needs to form a computer vision algorithm, some factors consider in advance.

1) Unknown subject category and unknown number of subject instances in video frame.
2) Complex motion of salient objects due to pose variation.
3) Ambiguous appearance in cluttered background.

It is infeasible to manipulate all foreground objects beforehand. One can extract foreground object from a video using foreground or background information. The extracted object can be utilised for further processing in video object recognition framework. Thus, the task of object recognition is done.

Besides the above approaches, graph based methods have been shown to be effective for foreground object segmentation. Using such methods, an image is typically represented by a graph, in which each observed node indicates an image pixel and the associated hidden node correspond to its label. By determining the cost between adjacent hidden nodes using color, motion, etc. information, one can segment the foreground object by dividing the graph into disjoint parts while minimizing the total cost. In this paper, we focus on VOE in single concept videos captured by a monocular camera in static or arbitrary types of background. Instead of assuming that the background motion is consistently dominant and different from that of the foreground (as [13] did), we relax this assumption and al low foreground objects to be present in scenes which have marginal but complex background motion (e.g., motion induced by sea waves, swaying trees, etc.). We also ignore the video frames with significant motion variations due to shot changes or abrupt camera movements. To make our method robust and not require any user interaction, we start from multiple local motion cues, and integrate the induced shape and color models into a CRF. As we will discuss in Section 2, shape features better preserve local information of the foreground object than motion cues do, and our proposed framework allows the use of both foreground and background color models to provide better generalization in formulating the associated CRF model. It is worth noting that, our method does not require the prior knowledge of the object category, and thus no training data or object part detectors are needed. All the feature models we utilize in our CRF are

DOI: 10.15680/IJIRSET.2014.0309063

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

automatically extracted from the test input video in an unsupervised setting, and this cannot be easily achieved by most prior work.

II. RELATED WORK

Previous work such as [8] and [9] focused on an interactive scheme and required users to manually provide the ground truth label information. Although excellent results were produced, methods which do not require user interaction are more practical for real world applications. Recently, several automatic segmentation techniques have been proposed. For example, Wu et al. [10] used a stereo camera setting which provides depth information as a cue for ground truth label. For videos captured by a monocular camera, literatures such as [11, 12] used a CRF framework which maximizes a joint probability of color, motion, etc. models to predict the label of each image pixel. Although the color features can be automatically determined from the input video, these methods required the trained object detectors to extract shape or motion features. A recently proposed method in [13] addressed the VOE problem without the use of any training data. It assumes that the motion of the background is dominant throughout the video, so the authors apply RANSAC [19] to extract candidate foreground regions, followed by a CRF which combines the associated color and motion features to determine the final foreground region.

On the other hand, unsupervised approaches do not train any specific object detectors or classifiers in advance. For videos captured by a static camera, extraction of foreground objects can be treated as a background subtraction problem. In other words, foreground objects can be detected simply by subtracting the current frame from a video sequence [24], [25]. However, if the background is consistently changing or is occluded by foreground objects, background modeling becomes a very challenging task. For such cases, researchers typically aim at learning the background model from the input video, and the foreground objects are considered as outliers to be detected. For example, an autoregression moving average model (ARMA) that estimates the intrinsic appearance of dynamic textures and regions was proposed in [26], and it particularly dealt with scenarios in which the back- ground consists of natural scenes like sea waves or trees.

Sun et al. [27] 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. For example, graph-based methods [28], [29] identify the foreground object regions by minimizing the cost between adjacent hidden nodes/pixels in terms of color, motion, etc. information.

More specifically, one can segment the foreground object by dividing a graph into disjoint parts whose total energy is minimized without using any training data. While impressive results were reported in [28], [29], these approaches typically assume that the background/camera motion is dominant across video frames. For general videos captured by freely moving cameras, these methods might not generalize well (as we verify later in experiments). Different from graph-based methods, Leordeanu and Collins [30] proposed to observe the co-occurrences of object features to identify the foreground objects in an unsupervised setting. Although promising results under pose, scale, occlusion, etc. variations were reported, their approach was only able to deal with rigid objects (like cars).

III. AUTOMATIC OBJECT MODELING AND EXTRACTION

Since not all the parts of a moving object will produce motion cues, or some of these cues might be negligible due to low contrast, etc. effects, it is not surprising that motion cues are not sufficient for VOE problems. To overcome this limitation, we propose to first extract the motion cues from the moving object across video frames, and we combine the motion induced shape, foreground and background color models into a CRF. Without prior knowledge of the object of interest, this CRF model is designed to address VOE problems in an unsupervised setting. In Section 2.1, we first briefly review the use of CRF for object segmentation/extraction. We will detail the construction of our motion, shape, foreground and background color models, and discuss how we integrate them into a unified CRF framework in the remaining of this section.

DOI: 10.15680/IJIRSET.2014.0309063

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

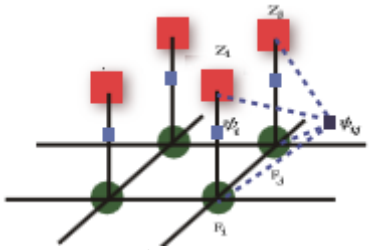

Fig 1: CRF for object segmentation

3.1. Extraction of visual saliency

To extract visual saliency of each frame, we perform image segmentation on each video frame and extract color and contrast information. In our work, we advance Turbopixels proposed by [21] for segmentation, and the resulting image segments (superpixels) are applied to perform saliency detection. The use of Turbopixels allows us to produce edge-preserving superpixels with similar sizes, which would achieve improved visual saliency results as verified later. For the kth superpixel r_k, we calculate its saliency score $S(r_k)$ as follows:

$$S(r_k) = \sum_{r_k \neq r_i} exp(D_s(r_k, r_i)/\sigma_s^2)\omega(r_i)D_r(r_k, r_i)$$
$$\approx \sum_{r_k \neq r_i} \exp(D_s(r_k, r_i)/\sigma_s^2)D_r(r_k, r_i) \quad (1)$$

where $D_s$is the Euclidean distance between the centroid of $r_k$and that of its surrounding superpixels r_i, while $\sigma_s$controls the width of the kernel. The parameter $\omega(r_i)$is the weight of the neighbor superpixel r_i, which is proportional to the number of pixels in r_i. Compared to [22], $\omega(r_i)$ can be treated as a constant for all superpixels due to the use of Turbopixels (with similar sizes). The last term $D_r(r_k, r_i)$ measures the color difference between $r_k$and r_i, which is also in terms of Euclidean distance. As suggested by [23], we consider the pixel i as a salient point if its saliency score satisfies $S(i)$>0.8∗max(S), and the collection of the resulting 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 $\hat{S}(i)$ for each pixel i as follows:

$$\hat{S}(i) = S(i) * (1 - dist(i)/dist_{max}) \quad (2)$$

where $S(i)$ is the original saliency score derived by (1), and dist(i) measures the nearest Euclidian distance to the salient point set. We note that distmax in (2) is determined as the maximum distance from a pixel of interest to its nearest salient point within an image, thus it is an image dependent constant.

3.2. Extraction of motion cues

In our work, each moving part of a foreground object is assumed to form a complete sampling of the entire object of interest (as [11, 12, 13] did). We aim to extract different feature information from these moving parts for the later CRF construction. To detect the moving parts and their corresponding pixels, we perform dense optical flow forward and backward propagation [15] at every frame. A moving pixel q_t at frame t is determined by:

$$q_t = \hat{q}_{t,t-1} \cap \hat{q}_{t,t+1} \quad (3)$$

Where q_t denotes the pixel pair detected by forward or backward optical flow propagation. Only if a pixel is identified by the opticalflow trajectories in both directions, we will denote it as a pixel of a moving object. To alleviate the influence of camera shake, we ignore the frames which result in a large number of moving pixels after this step. After determining the regions induced by the moving object (or its parts), we will extract the associated shape and color information from these regions, as we discuss next.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

3.3. Learning shape cues

Since we assume each moving part of an object forms a complete sampling of the entire object, part based shape information induced by the above motion cues can be advanced to characterize the foreground object. To describe each moving part, we apply the histogram of oriented gradients (HOG) features. We first divide each frame into disjoint 8×8 pixel grids, and we compute HOG descriptors for each region (patch) of 4×4 = 16 grids. To capture scale invariant shape information, we further downgrade the resolution of each frame and repeat the above process (the lowest resolution of the scaled image is a quarter of that of the original one). We note that [20] also used a similar setting to extract their HOG descriptors. Since the use of sparse representation has been shown to be very effective in many computer vision tasks [16], once the HOG descriptors of the moving foreground regions are extracted, we learn an over complete codebook and determine the associated sparse representation of each HOG.

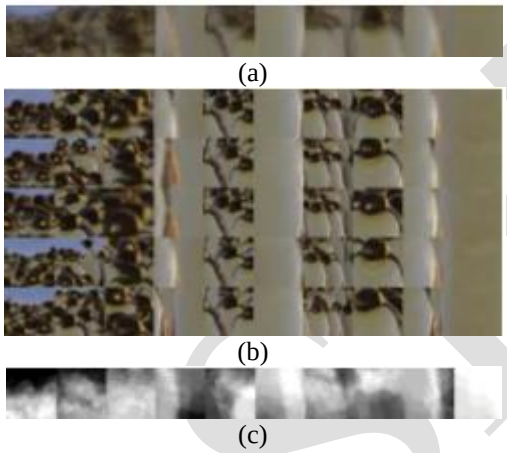

Fig 2: Visualization of sparse shape representation. (a) Example codewords for sparse shape representation. (b) Corresponding image patches (only top 5 matches shown). (c) Corresponding masks for each codeword.

After obtaining the dictionary and the masks to represent the shape of foreground objet, we use them to encode all image patches at each frame. This is to recover non moving regions of the foreground object which does not have significant motion and thus cannot be detected by motion cues. For each image patch, we derive its sparse coefficient vector, and each entry of this vector indicates the contribution of each shape codeword. Correspondingly, we use the associated masks and their weight coefficients to calculate the final mask for each image patch. The reconstruction image using foreground shape information is then formulated as:

$$\hat{X}_t^S = \sum_{n \epsilon I_t} \sum_{k=1}^{K} \alpha_{n,k} \cdot M_k \qquad (4)$$

Figure 3 shows an example of the reconstruction of a video frame using shape information of the foreground object (induced by motion cues only). We note that b XS t serves as the likelihood of foreground object at frame t in terms of shape information. This shape likelihood function contributes to the shape energy function in CRF, i.e.

$$E^S = -w^s \log(\hat{X}_t^S) \qquad (5)$$

where w^s controls the contribution of this shape energy term in the final CRF formulation. Comparing to the motion likelihood in Section 2.2 and [13], it is expected that better candidate foreground object can be discovered using the above motion induced shape information. This makes the use of foreground and background color models more feasible, as we discuss next.

DOI: 10.15680/IJIRSET.2014.0309063

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

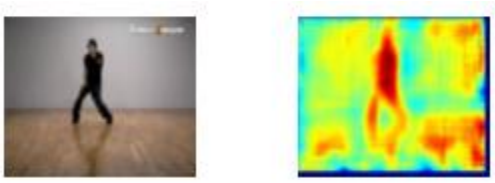

Fig 3: Shape Likelihood

3.4. Learning color cues

Besides the motion induced shape information, we also combine the color cues into our CRF framework to better model the object of interest.

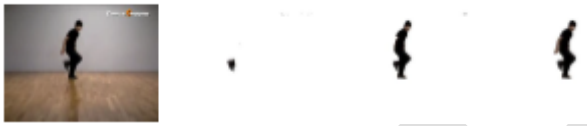

Fig 4: Object extraction example of (a) the input frame, using (b) motion, (c) foreground color, and (d) our proposed CRF integrating multiple types of motion induced features.

Constructing the background model is difficult because the sparse motion cues throughout the video might not be sufficient to indicate the foreground/background regions. This difficulty can be depicted in Figure 4(b) which is an example of foreground region extraction using only motion cues in a CRF. To apply the color information of both foreground objects and the remaining background regions into our CRF, we utilize the shape likelihood image obtained from the previous step, and threshold the resulting shape posterior probability $\hat{X}^S$. For the pixels of $\hat{X}^S$ whose probability values are above a predetermined threshold, the associated regions will be potentially considered as foreground; those below the threshold will be thus grouped as candidate background regions. For these candidate foreground and background regions, we use Gaussian Mixture Models (GMM) G^{cf} and G^{cb} to model the RGB distribution for each, with the number of Gaussian components set to 10 for both cases. We now detail the learning of color cues for foreground object extraction. A single energy term which is associated with both foreground and background color models in our CRF is defined as follows:

$$E^c = E^{cf} - E^{cb} \qquad (6)$$

Where

$$\begin{cases} E^{CF} = -w^{c_f} \log(\sum_{i \epsilon I} G^{Cf}(i)) \\ E^{cb} \\ E^{CB} = -w^{c_b} \log(\sum_{i \epsilon I} G^{Cb}(i)) \end{cases}$$

As for the visual saliency cue at frame t, we convert the visual saliency score $\hat{S}_t$ derived in (2) into the following energy term E^V :

$$E^V = -w^v \log(\hat{S}_t) \qquad (7)$$

Similar to (6), w^{c_f} and w^{c_b} in (7) weight the corresponding color models in the CRF. It is worth noting that only foreground color information is modeled in [13]. As we will show later in our experiments, disregard of the background color model would limit the performance of object segmentation.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

3.5. Integration of multiple feature models via CRF

Induced by motion cues, we combine both shape and color (foreground and background) models into our CRF framework. Since we do not require prior knowledge of the object category, use of multiple types of motion induced features allows us to model the foreground object of interest without the need of user interaction or any training data. To provide the property of spatial coherence into our CRF model, we introduce a pairwise term to preserve local foreground/background structures.

We note that the above pairwise term is able to produce coherent labeling results even under low contrast or blurring effects. Finally, by integrating (6), (7), the objective energy function (2) and pair wise term of our CRF can be re written as:

$$\begin{aligned} E &= E_{unary} + E_{pairwise} \\ &= (E^S + E^{CF} - E^{CB}) + E_{i,j} \\ &= E^S + E^C + E^V + E_{i,j} \end{aligned} \qquad (8)$$

To solve the above optimization problem, one can apply graph based energy minimization techniques such as max flow/min cut algorithms. When the above energy function is minimized, the labeling function output F indicates the class label (foreground or background) of each observed pixel.

IV. FOREGROUND OBJECT RECOGNITION

Object recognition is a process for identifying a specific object in a digital image or video. Here, neural networks are used for object recognition purpose. Neural Network provides functions and apps for modeling complex nonlinear systems that are not easily modeled with a closed form equation. Neural Network supports supervised learning with self organizing maps and competitive layers. With this we can design, train, visualize, and simulate neural networks.

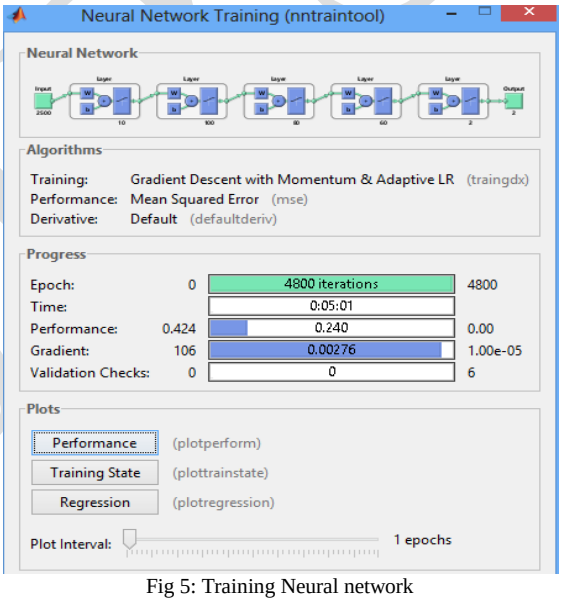

Fig 5: Training Neural network

V. CONCLUSION

In this paper, we proposed a method which utilizes multiple motion induced features such as shape and fore ground/background color models to extract foreground objects in single concept videos. We advanced a unified CRF framework to integrate the above feature models. Using sparse representation techniques, our motion induced shape

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

model describes the shape information of the foreground object in a probabilistic way, which allows us to extract and construct both foreground and background color models for the object of interest. Compared with prior work, our approach better models the foreground object due to the use of multiple types of motion induced feature models, while no prior knowledge of the object of interest, collection of training video data, or the design of object part detectors are required.

Future research will be directed at extensions of our approach for videos with multiple concepts (i.e. multiple foreground objects of interest), and the applications of VOE for higher level tasks such as action/activity recognition and video retrieval. For these applications, we expect to integrate features from heterogeneous domains (e.g. visual, audio, temporal, text, etc.), and we will provide a systematic way to select proper feature models for extracting particular types of the object of interest.

REFERENCES

[1] T. Bouwmans et al., "Background modeling using mixture of gaussians for foreground detection – a survey," Recent Patents on Computer Science, 2008.
[2] F. C. Cheng, S. C. Huang, and S. J. Ruan, "Advanced background subtraction approach using laplacian distri bution model," in ICME, 2010.
[3] B. Wu and R. Nevatia, "Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses," IJCV, 2009.
[4] Z. Lin and L. S. Davis, "Shape based human detection and segmentation via hierarchical part template matching," IEEE PAMI, 2010.
[5] L. Gorelick et al., "Shape based detection and top down delineation using image segments," IEEE PAMI, 2009.
[6] J. C. Niebles et al., "Extracting moving people from internet videos," in ECCV, 2008.
[7] J. C. Niebles, B. Han, and L. Fei Fei, "Efficient extraction of human motion volumes by tracking," in CVPR, 2010.
[8] Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images," in ICCV, 2001.
[9] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut": Interactive foreground extraction using iterated graph cuts," ACM Trans. Graphics, 2004.
[10] Y. Wu et al., "Bilayer segmentation from stereo video sequences by fusing multiple cues," in ICME, 2008.
[11] A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov, "Bilayer segmentation of live video," in CVPR, 2006.
[12] P. Yin et al., "Bilayer segmentation of webcam videos using tree based classifiers," IEEE PAMI, 2010.
[13] F. Liu and M. Gleicher, "Learning color and locality cues for moving object detection and segmentation," in CVPR, 2009.
[14] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in ICML, 2001.
[15] M. Werlberger et al., "Anisotropic Huber L1 optical flow," in BMVC, 2009.
[16] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," IEEE Trans. Image Processing, 2006.
[17] J. Mairal et al., "Online learning for matrix factorization and sparse coding," JMLR, 2010.
[18] A. Blake et al., "Interactive image segmentation using an adaptive gmmrf model," in ECCV, 2004.
[19] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," in Commun. ACM, 1981.
[20] P. Felzenszwalb, et al., "Object detection with discriminatively trained part based models," IEEE PAMI, 2010.
[21] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, and S. J. Dickin- son, "TurboPixels: Fast superpixels using geometric flows," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 12, pp. 2290–2297, Dec. 2009.
[22] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2011, pp. 409–416.
[23] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 2376–2383.
[24] [9] T. Bouwmans, F. E. Baf, and B. Vachon, "Background modeling using mixture of Gaussians for foreground detection—A survey," Recent Patents Comput. Sci., vol. 3, no. 3, pp. 219–237, 2008.
[25] F.-C. Cheng, S.-C. Huang, and S.-J. Ruan, "Advanced background subtraction approach using Laplacian distribution model," in Proc. IEEE Int. Conf. Multimedia Expo, Jul. 2010, pp. 754–759.
[26] J. Zhong and S. Sclaroff, "Segmenting foreground objects from a dynamic textured background via a robust Kalman filter," in Proc. 9th IEEE Int. Conf. Comput. Vis., vol. 1. Oct. 2003, pp. 44–50.
[27] J. Sun, W. Zhang, X. Tang, and H.-Y. Shum, "Background cut," in Proc. 9th Eur. Conf. Comput. Vis., 2006, pp. 628–641.
[28] F. Liu and M. Gleicher, "Learning color and locality cues for moving object detection and segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 320–327.
[29] K.-C. Lien and Y.-C. F. Wang, "Automatic object extraction in single- concept videos," in Proc. IEEE Int. Conf. Multimedia Expo, Jul. 2011, pp. 1–6.
[30] M. Leordeanu and R. Collins, "Unsupervised learning of object features from video sequences," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2005, pp. 1142–1149.