Motion Characterization of a Dynamic Scene

Keywords: Image and Video Analysis, Scene Understanding, Segmentation and Grouping

Abstract: Given a video, there are many algorithms to separate static and dynamic objects present in the scene. The proposed work is focused on classifying the dynamic objects further as having either repetitive or non-repetitive motion. In this work, we propose a novel approach to achieve this challenging task by processing the optical flow fields corresponding to the video frames of a dynamic natural scene. We design an unsupervised learning algorithm which uses functions of the flow vectors to design the feature vector. The proposed algorithm is shown to be effective in classifying a scene into static, repetitive, and non-repetitive regions. The proposed approach finds significance in various vision and computational photography tasks such as video editing, video synopsis, and motion magnification.

1 Introduction

Computer vision involves estimation of scene information, which the human vision can perceive very easily, from images and videos using efficient algorithms. One of the challenging problems in computer vision is the identification of the type of motion a given object exhibits in a natural scene. Analysing motions of different objects in a scene might be a trivial task for a human being. However, it is extremely complex for a computer. This complexity is due to the differences in the appearance of the objects and the different types of motion each object may undergo at a given time. Given a digital image/video, computer vision researchers strive to perform high level vision tasks such as recognition and segmentation.

However, the digital video captured is just a 2D projection of the 3D scene being captured (Peterson, 2010). A set of consecutive video frames provide necessary information for the segmentation of a scene depending on the types of motion present. A scene may have static, repetitive, and non-repetitive motion regions. Algorithms based on optical flow yield flow fields that form the basis for designing feature vector for each pixel location. For a general natural scene, displacement flow vectors of objects could exhibit a wide range of variations. Hence, segmenting such scenes is a major challenge. Segmentation of different motion regions in a dynamic scene has various applications such as removal of occlusion, scene categorization and understanding, video editing, video synopsis, motion magnification, to name a few.

Sampled video frames of a scene having a rotating wheel are shown in Fig. 1(a). Some of the frames extracted from the video are shown in Fig. 1(b). The extracted frames indicate the presence of repetitive object in the scene (wheel) and also show the presence of non-repetitive motion in an object (hand). The scene also has static regions. Fig. 1(c), 1(d), and 1(e) depict the various types of motion regions present in this scene. The presence of non-repetitive motion (like the one shown in Fig. 1(e)) is common in real world scenes. Our approach aims at classifying these different types of motion automatically given a video corresponding to any natural scene.

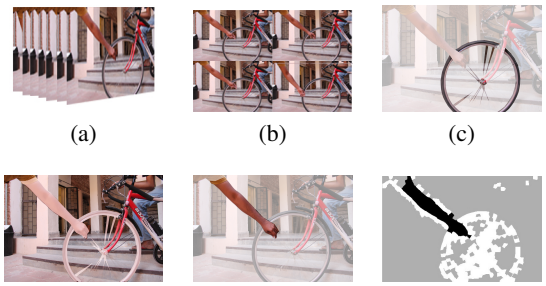


Figure 1: (a) Video of a dynamic scene, (b) the frames of a video corresponding to a dynamic scene, (c), (d), and (e) illustrate the repetitive, static, and non-repetitive regions of the scene, and (f) output segmentation from the proposed algorithm.

We address this problem by designing a novel segmentation algorithm. We shall design robust feature vectors to separate repetitive, non-repetitive, and static regions in a natural scene. This classification would enable us to categorise these regions and use them to solve other computer vision applications. Our approach can also find application in compressing the videos. Thus we have a diverse set of applications of

the proposed approach in vision, computational photography, and video processing. For instance, redundant multiple times recording of repetitive motion in a scene such as rotating fan motion, flowing river, periodic sea waves and others can be avoided by proper segmentation of repetitive part of the scene from the video. This is the one of the fruitful application of our proposed approach for three label segmentation.

The primary contributions of the proposed work are:

1. Design of a novel feature descriptor to classify the static, repetitive, and non-repetitive motion regions in a scene.
2. Design of an unsupervised learning framework for bottom-up segmentation of these regions.
3. The feature vectors are modelled as functions of the contents from space time volume with finite time support for efficient performance.
4. The approach does not depend on object surface properties such as reflectance, texture, color, etc. and predicts the type of motion an object undergoes even for objects with different appearances.

In Section 2, a brief description about the related research work is provided. Section 3 contains a description about the design of feature descriptor for unsupervised learning. In Section 4, results of applying the proposed feature vector on several dynamic scene data sets are described. Section 5 presents the challenges facing the proposed algorithm. Section 6 provides directions regarding future work and section 7 provides the conclusion of the present work.

2 Related Work

Lucas-Kanade (Lucas and Kanade, 1981) and Horn-Schunck (Horn and Schunck, 1981) are the standard optical flow algorithms that mostly act as building blocks to separate static and dynamic objects in a scene. These algorithms operate efficiently under the assumption that the objects undergo small displacements in successive frames. However even if the above assumption is satisfied, the algorithm may not give good results for scenes which violate brightness constancy assumption and scenes which have transparency, depth discontinuity, and specular reflections. Some of these shortcomings have been overcome in the approach which uses feature descriptor matching (Black and Anandan, 1996). In order to get more efficient results in situations involving large object displacements, a more recent work on optical flow uses a hierarchical region based matching (Brox and Malik, 2011).

The flow fields obtained from an optical flow algorithm serve as the basic ingredients in the recent algorithms developed in the domain of video processing and computational photography. To extend its usefulness to situations involving large displacements, the Large Displacement Optical Flow (LDOF) algorithm was developed (Brox and Malik, 2011). This algorithm also uses a feature descriptor. An additional constraint in the energy optimization equation along with other constraints are used to establish the matching criteria in LDOF.

There are motion based segmentation algorithms that aid in background subtraction (Stauffer and Grimson, 1999). These approaches create a Gaussian mixture model for each pixel and segment static and dynamic pixels using a threshold. The existing algorithms in video synopsis rely on the extraction of dynamic objects in the scene for video synopsis (Pritch et al., 2008). They extract these interesting objects from the video and store them along with the timing information.The video is then condensed in which only the interesting objects are shown with their timing information.

The recent work on editing small movements using motion analysis of image pyramids involves the study of phase variations of motion dependent parameters. The phase variations are processed temporally to remove or enhance minute changes over time. The processing does not involve optical flow estimation and is therefore suitable for processing of videos of scenes where optical flow based approaches may fail (Wadhwa et al., 2013).

One of the recent works in the domain of moving object segmentation is given by Bergh and Van Gool (den Bergh and Gool, 2012). Their paper presents a novel approach of combining color, depth and motion information (optical flow) for segmenting the object using superpixels. This approach takes into consideration the 3D position of object and its direction of motion while segmenting the object. Another notable work on motion segmentation is given by Ochs and Brox (Ochs and Brox, 2012). Their paper presents an efficient approach of separating the moving objects from a video using spectral clustering.

3 Proposed Approach

We use two of our own datasets of dynamic scene each containing a set of seven images captured in the burst mode with a DSLR camera. Additionally, we use YUPenn[1] dataset of dynamic scenes consisting of

[1] http://www.cse.yorku.ca/vision/research/dynamic-scenes/

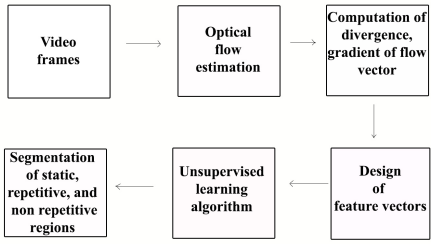


Figure 2: The Proposed Approach

14 different scenes (Derpanis and Wildes, 2012). We sampled a set of 5-7 images from each dynamic scene video present in the YUPenn dataset. In the scenes, we come across objects that exhibit a large displacement for which the standard algorithms such as Lucas Kanade flow vector usually fail. Therefore, we use the velocity vectors obtained from the LDOF algorithm in these cases. For the motions with relatively small displacement compared to the object size, Lucas-Kanade algorithm is sufficient to achieve good results. Usage of appropriate flow vectors are done manually depending on the type of motion- small or large displacement movements. The Optical flow algorithms applied on every two consecutive images gives 2-D vector field consisting of optical flow vectors.

We consider the first image as the reference image as optical flow vector of a frame is computed with reference to another frame. We use the optical flow vectors for the segmentation of the reference image into different motion regions corresponding to the natural scene. We use divergence of the flow vector and the gradient of magnitude of the flow vector of the image at every point in the image to construct the feature vector. The divergence of a vector field is the rate at which flux exits a given region of space. Gradient represents the magnitude and direction of maximum variation of the scalar field (image).

Let

$$P_i(x,y) = (\overrightarrow{V_x}, \overrightarrow{V_y}, \overrightarrow{\nabla}.\overrightarrow{V}, \overrightarrow{\nabla}_x|\overrightarrow{V}|, \overrightarrow{\nabla}_y|\overrightarrow{V}|) \quad (1)$$

$$P_i(x,y) = (p_1, p_2, p_3, p_4, p_5) \quad (2)$$

where $P_i(x,y)$ is a vector at a given pixel location (x,y) for i^{th} frame.

Here p_1 and p_2 are the optical flow vectors at *(x,y)*. After applying the vector operations on the flow vectors, we obtain their divergence and gradient of flow vectors for every pixel in the image. p_3 corresponds to the divergence and p_4 and p_5 represent the gradients of magnitudes of the vectors of the given frame (image) as shown in Equation 1.

For better classification, we need to add finite temporal support using information from successive images. Consequently, we take about 5-6 frames from a part of a 5 second video. We take the variance of divergence of flow vectors, and gradient of magnitude of flow vector, as in Equation 1, on each pixel across the video frames in temporal window. An intuitive reason behind using variance of divergence and gradient is justified by the difference in variations in their values for different regions. For static regions, the variation will be negligible because the magnitude of velocity vectors in these regions will be close to zero over the temporal support of about 5 seconds as in the example shown in the Fig. 3. Non-repetitive motion regions will have a high variation due to their continuous variation in movement and repetitive motion regions will have a continuous but a smaller variation in their velocity vectors.

Let Q be the feature vector that is used in the unsupervised learning algorithm. There is a significance behind bringing Q vector in Equation 3 into logarithmic domain . The formation of margin for three region clustering takes place better in logarithmic domain as it can be clearly seen from Fig. 3(b) and 3(f).

Let

$$Q_i(x,y) = (q_1, q_2, q_3, q_4, q_5), \quad (3)$$

where $q_j = \log(a + \sigma^2(p_j))$, $j = 1,2,3,4,5$,

where σ^2 is the variance of a feature vector across the space-time volume over finite time support, and a is a very small non-zero, positive constant (a=0.1 for Fig. 3). We include this constant to avoid condition when the variance becomes close to zero (especially in static regions) as the logarithm value may tend to infinity. This parameter is purely experimental, we varied its value from 0.01 to 1 for different cases in the Fig. 4 depending on the corresponding segmentation output.

The number of pixels in a low resolution image are high which make optimization at pixel level quite difficult. Superpixels group a set of neighbouring pixels which share similar properties. Using superpixels, we preserve the boundaries of the objects in the image and it will help in reducing the effective number of pixels in the image (Ren and Malik, 2003). This increases the calculation efficiency during the execution of program for motion classification in a scene. The superpixel based approach improves accuracy in classification by finding features for each superpixel rather than for individual pixels. When we plot the divergence of optical flow vector field, we may get noisy results as observed in Fig. 3(c). Superpixels help in the suppression of noise as we take the average of the feature vectors estimated within a superpixel and this operation helps in improving the accuracy while performing segmentation of the scene.

Every pixel within a superpixel is uniformly as-

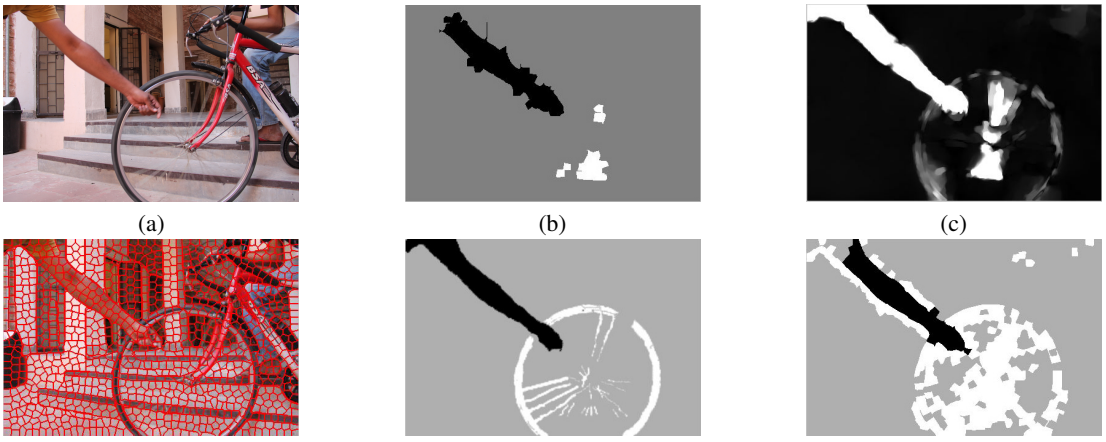


Figure 3: (a) One among the input frame, (b) 3 class Segmentation using K-means clustering using V_x and V_y of $\overrightarrow{V}$ in Equation 1, (c) using LDOF on pixels lead to noisy segmentation, (d) result of applying superpixel on reference image, (e) expected result (ground truth), (f) segmentation result using the proposed algorithm. (white - repetitive, black - non-repetitive, gray - static).

signed a particular feature vector. The individual elements are calculated as the mean of the feature vector values corresponding to the pixels present in a superpixel. This feature vector is used in the unsupervised learning algorithm that clusters the feature points into three different classes. In our experiment, we use K-means clustering algorithm for testing the designed feature vector. This enables us in the segmentation of the scene according to the reference image.

Addition of more information to the feature vector by including divergence for each set of flow vector obtained from more frames improves the result of classification. Feature vector comprises of function of variance of optical flow vectors, divergence of flow vectors, and gradients of its magnitude. This make the flow vector into 5-dimensional motion flow vector(5DMFV). Optical flow vectors obtained from different methods are compared by analysing the resultant classification with the ground truth image created for the scene (see Fig. 3). Applying K-means clustering for 3 clusters on this 5DMFV categorises into 3 classes- static, repetitive and non-repetitive motion, if present in the scene. As seen in Fig. 3, certain scenes have significant variation in motion from static and small regular motion to random motion. This is the reason that interests us to group the entire 5DMFV space into 3 clusters.

4 Results

Fig. 5 depicts the comparison of our approach with one of the recent works in motion segmentation

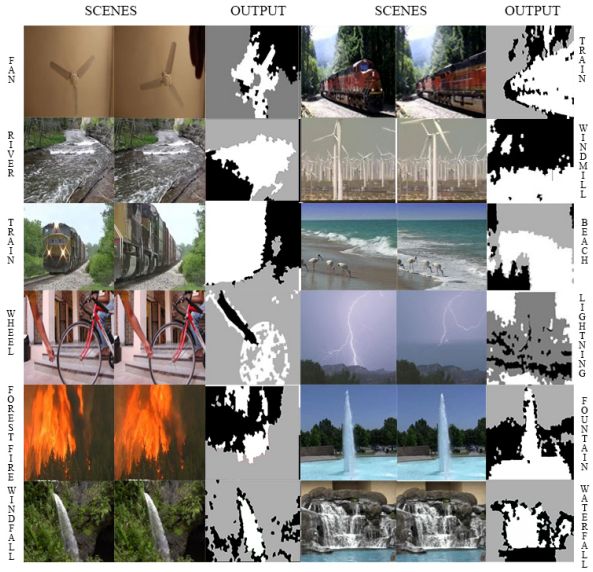


Figure 4: Various datasets and segmentation results (white - repetitive, black - non-repetitive, gray - static)

based on point trajectories. Column-1 of the Fig. 5 has a frame extracted from a video of a real world scene. Figures in column 2 and 3 are the results obtained from the binaries[2] of motion segmentation approach of Ochs *et al*. We have presented the results of our proposed approach with just optical flow vec-

[2]http://lmb.informatik.uni-freiburg.de/resources/software.php

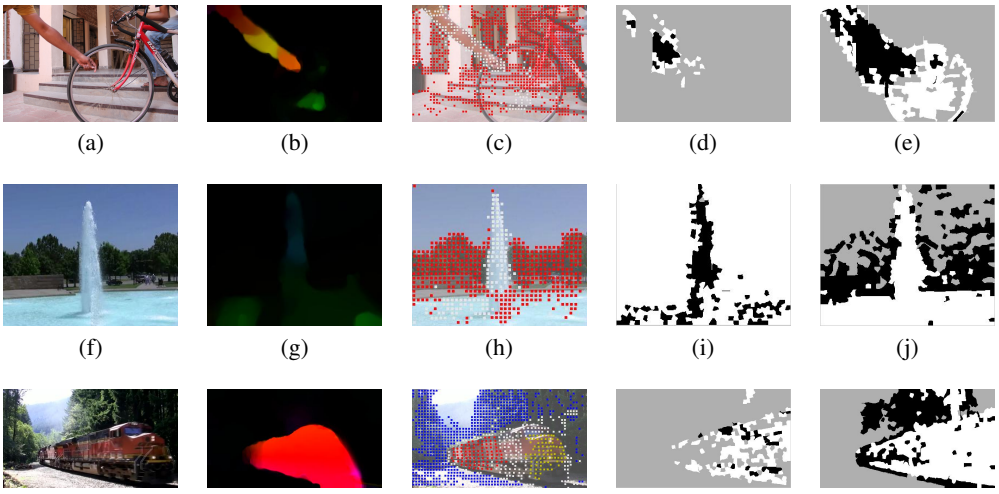


Figure 5: (a, f, k) Real video sequence of Wheel, Fountain and Train, (b, g, l) Result obtained from binaries from (Ochs and Brox, 2012), (c, h, m) Trajectory clustering on video sequence from Ochs *et al* (Ochs and Brox, 2012), (d, i, n) Results of proposed approach using Lucas and Kanade optical flow vectors as feature vectors, (e, j, o) Proposed Approach result using 5-dimensional feature vector. Here, white depicts repetitive motion, black depicts non-repetitive motion and gray indicates static region in the scene

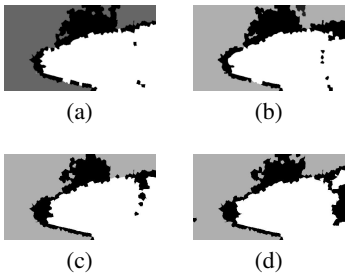
tor as feature vector in the column-4 while column-5 shows results obtained using 5DMFV. In these scenes, we have different kinds of motion like rotational motion of wheel, moving hand, train, fountain, trains and among others.

As seen from the first row, the reference image has static background, repetitive motion of wheel and non-repetitive motion of hand. Although (Ochs and Brox, 2012) approach is able to segment the hand clearly, the wheel segmentation is not perfect. Fig. 5(d) shows that the segmentation result leads to many errors. Fig. 5(e) depicts the results of the proposed approach which is able to segment the reference image into three regions comprising of static, repetitive and non-repetitive motions. Similarly, we have the fountain scene with repetitive fountain and non repetitive far off trees movement. Also we consider the railway scene with constant locomotive motion and irregular tree movement due to the emitted smoke. In the fountain scene, the work by (Ochs and Brox, 2012) seems to perform better with less erroneous regions in the trajectory clustering image when compared to Fig. 5(j) which segmented three parts with some errors. However, proposed approach makes appropriate segmentation in the railway scene with white as repetitive locomotive motion, black as tree movement and gray as the static region as shown in Fig. 5(o) in comparison to Fig. 5(l, m). We conclude from Fig. 5 that the proposed approach leads to promising results. Thus, we bring out a visual comparison of the above methods. Their quantitative comparison is not possible because state of the art method deals with segmentation of static and dynamic parts of a scene.

We apply our algorithm to a set of five consecutive frames of the wheel scene. In Fig. 3, the wheel exhibits repetitive motion while the arm exhibits non repetitive motion. Frames of the video used are shown in Fig. 3(a) and 3(b). We use the first image in our set as the reference image. Fig. 3(c) shows that the segmentation at pixel level is noisy. In the reference image, super pixels are calculated as shown in Fig. 3(d) and these superpixels are employed for the segmentation. Upon application of our algorithm, the result shown in Fig. 3(f) is obtained which matches closely with the approximate ground truth image (Fig. 3(e)). Fig. 4 depicts two representative images from each dataset that were used and their corresponding outputs.

Fig. 6 shows the segmentation of railway scene in Fig. 5(k) at different sampling rates and different number of frames used. From the figure shown, it is seen that output of segmentation is invariant of both sampling rate and number of frames used for designing feature vector. Thus the proposed feature vector design is robust to sampling rate and number of frames used.

The error plot for various sampling values estimated for the train dataset (Fig. 7). The error represents the percentage of mis-matched pixels between ground truth image and output of proposed approach,



(a) (b) (c) (d)

Figure 6: Segmentation of railway scene shown at sampling rates of, (a) 1 per 15 frames with a total of 10 frames, (b) 1 per 14 frames with a total of 11 frames, (c) 1 per 13 frames with a total of 12 frames, (d) 1 per 12 frames with a total of 13 frames. Gray - Static, White - Repetitive and Black- Non-repetitive

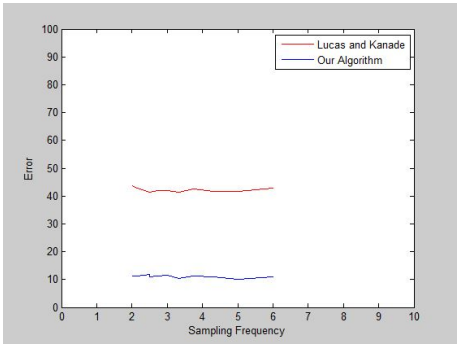
more precisely Fig. 7 depicts error between Fig. 3(e) and 3(f). The error value estimated when used the proposed feature vector, is compared against using just the estimated flow vectors V_x and V_y of $\overrightarrow{V}$ in Equation 1 as feature vectors. We can observe vast improvement in error performance when the proposed feature vectors are used. In addition to that, we can infer that error is independent of the sampling frequency of extraction of frames from the video. This shows the robustness of the proposed approach with respect to the sampling frequency.

When we apply the proposed algorithm on the river dataset (Fig. 8), we observe that the distant part of the river is classified as static, nearer part is classified as non-repetitive and the middle part is classified as repetitive. For a pinhole camera, the 2D perspective projection matrix of the camera is provided in Equation 4.

$$\underbrace{\begin{bmatrix} x \\ y \\ 1 \end{bmatrix}}_{\text{Projected point}} = \underbrace{\begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{\text{Projective Matrix}} \underbrace{\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}}_{\text{World co-ordinates}} \quad (4)$$

where f is the distance of optical center from image plane. Z is the depth of the scene. After projection, the point whose 3D co-ordinates are (X,Y,Z) is projected on image plane as $(fX/Z, fY/Z)$.So when object is far, depth variations are not felt and the projection becomes orthographic projection. In this case, the 3D scene point is projected as (X,Y).

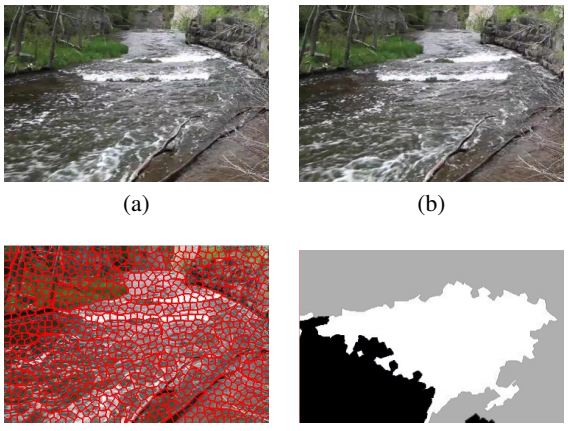
When object is near, depth variation is appreciable and the projection becomes strong perspective projection. In cases where the depth is slightly more, the projection becomes weak perspective projection. Therefore the variation in depth within the same object may result in different classifications of its parts.



(a)

Figure 7: Error performance comparison of the proposed feature vector as against flow vectors as feature vector for the train dataset

Fig. 8(d) shows that the distant region of the river is classified as static region while the nearer region of the river is classified as having non-repetitive motion and rest of the region is classified as having repetitive motion.



(a) (b) (c) (d)

Figure 8: (a) and (b) Consecutive frames of river scene. (c) the result of applying superpixel on reference image, and (d) segmentation result using the proposed algorithm (white - repetitive, black - non-repetitive, gray - static).

5 Challenges

There may be scenes that have very large object displacements where both normal optical flow method and LDOF algorithm yield less accurate results. Though the scene may contain regions of non-repetitive motion, they may get unidentified because of its absence in the sampled images. The algorithm

may fail, for instance, when there is a lightning in the scene. This is expected as optical flow algorithm works under the assumption of constant brightness.

Change of lighting condition in the scene leads to error in the segmentation. Optical flow algorithms have its dependency on the brightness value at the pixel location. Segmentation problems arise in such exceptional cases of complex natural scenes. When the depth of a moving object varies widely, then different regions of the same object will be classified as exhibiting different motion even though the object as a whole exhibits same motion when observed with eyes (See Fig. 8). Usage of unsupervised learning such as K-means clustering gives rise to the problem of different partitions resulting in different clusters.

Mathematical representation of repetitive or regular motion part of a scene is difficult for a real world scene. Repetitiveness varies in each scene. Our approach is formulated for any general scene with well defined variation in motion.

6 Future Work

Classification of motion in a dynamic scene has a bright research future when the scene is affected by drastic illumination changes. In some of the previously considered examples, we saw that the illumination of the scene keeps fluctuating which leads to bad results upon implementation of the proposed algorithm (Fig. 4). There may be problems due to variations in camera parameters such as aperture, focal length, and shutter speed. We plan to improve the proposed approach for use in video synopsis and motion magnification in future. Another future work can be scene classification and categorization by the usage of supervised learning, if we provide training examples for these motion categories. Building a dataset for the different motion categories corresponding to the natural scene is a work in progress.

7 Conclusion

The proposed approach segments the scene in a finite timed video (of about 5 seconds) into static, repetitive, and non-repetitive motion regions effective for a sampling rate between 1 per 30 frames to 1 per 5 frames. For scenes containing large displacements, LDOF gives better results. Optical flow algorithms such as Lucas-Kanade and LDOF alone do not classify the motion while implementation of special functions on the flow vectors produce better results. The approach fails in the scenes where lighting condition changes as the brightness constancy assumption does not hold true. Also when the depth of the object varies widely, we face difficulty in classification. We hope to customize this approach to other computer vision applications involving segmentation of different objects based on the motion they exhibit.

REFERENCES

Black, M. J. and Anandan, P. (1996). The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75 – 104.

Brox, T. and Malik, J. (2011). Large displacement optical flow: descriptor matching in variational motion estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(3):500–513.

den Bergh, M. V. and Gool, L. J. V. (2012). Real-time stereo and flow-based video segmentation with superpixels. In *WACV*, pages 89–96. IEEE.

Derpanis, K. G. and Wildes, R. (2012). Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(6):1193–1205.

Horn, B. K. and Schunck, B. G. (1981). Determining optical flow. *Artificial intelligence*, 17(1):185–203.

Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision (ijcai). In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81)*, pages 674–679.

Ochs, P. and Brox, T. (2012). Higher order motion models and spectral clustering. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 614–621. IEEE.

Peterson, B. (2010). *Understanding Exposure: How to Shoot Great Photographs with Any Camera*. Amphoto Books.

Pritch, Y., Rav-Acha, A., and Peleg, S. (2008). Nonchronological video synopsis and indexing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(11):1971–1984.

Ren, X. and Malik, J. (2003). Learning a classification model for segmentation. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 10–17 vol.1.

Stauffer, C. and Grimson, W. E. L. (1999). Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2. IEEE.

Wadhwa, N., Rubinstein, M., Durand, F., and Freeman, W. T. (2013). Phase-based video motion processing. *ACM Trans. Graph. (Proceedings SIGGRAPH 2013)*, 32(4).