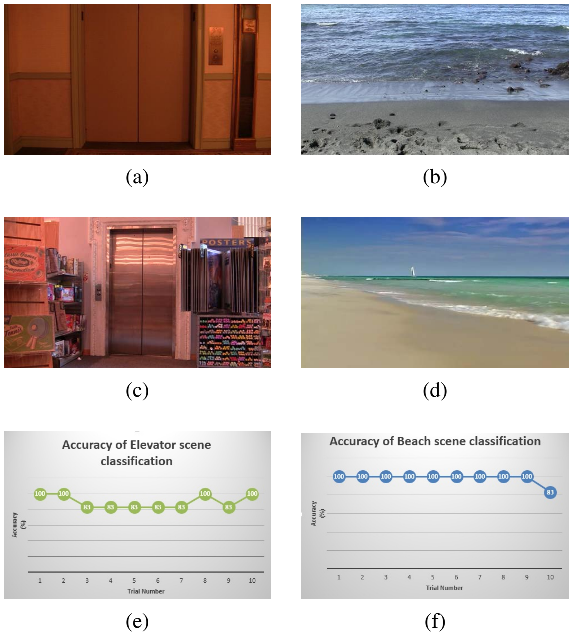
Dynamic Scene Classification using Spatial and Temporal Cues

Arun Balajee Vasudevan, Srikanth Muralidharan,
Shiva Pratheek Chintapalli
Indian Institute of Technology Jodhpur
Rajasthan, India
{arunbalajeev, srikanth, shiva}@iitj.ac.in

Shanmuganathan Raman
Indian Institute of Technology Gandhinagar
Gujarat, India
shanmuga@iitgn.ac.in

Abstract

A real world scene may contain several objects with different spatial and temporal characteristics. This paper proposes a novel method for the classification of natural scenes by processing both spatial and temporal information from the video. For extracting the spatial characteristics, we build spatial pyramids using the spatial pyramid matching (SPM) algorithm on SIFT descriptors while for the motion characteristics, we introduce a five dimensional feature vector extracted from the optical flow field. We employ SPM on combined SIFT and motion feature descriptors to perform classification. We demonstrate that the proposed approach shows significant improvement in scene classification as compared to the SPM algorithm on SIFT spatial feature descriptors alone.



(a) (b)

(c) (d)

(e) (f)

Figure 1: (a) and (c) are the video frames from two different elevator scenes, (b) and (d) are the frames from two different videos of beaches, (e) and (f) show the accuracy in classification of elevator and beach scenes over 10 trials respectively.

1. Introduction

An object present in a natural scene can be classified using its spatial as well as temporal characteristics. This task is easier for the human visual system (HVS). But this kind of semantic category identification of the objects present in a scene is very challenging and complex for the computer due to changes in appearance, illumination, view point, imaging system settings, etc.

A majority of existing scene classification algorithms use the spatial information for identifying the scene category. On the other hand, we take into account temporal motion in the natural scenes such as waves in beach scenes, snowfall on a windy night, and rotation of the wheels in a distant windmill. We extract this useful information from the temporal support in order to enhance the scene classification task.

There are several algorithms existing in the literature that use spatial low and high level image representations. The most famous low level image representation based algorithms use GIST [22] and histogram of oriented gradients (HOG) [21] for the classification task. The main problem with descriptors like HOG is that they are not robust enough for generalized tasks like scene classification and hence are limited to specific tasks like action recognition. Some of the high level image representation is useful in scene and object recognition including action recognition [23] and single image object classification [17]. All these approaches use information from the spatial domain giving little significance to temporal support for achieving the classification task.

This paper aims to make use of both spatial as well as temporal support obtained from the the video frames to identify the semantic category of the object present in the scene. The proposed approach is also different from other spatio-temporal based scene classification approaches. Dynamic scene classification in [6] is done by analysing orientation measurements related to spatial and temporal domains at multiple scales, where as in the proposed approach, we identify and extract feature vector from the dynamic regions of video scene. The proposed approach shows that by incorporating the motion information along with spatial information (SIFT features) for scene classification, there is a significant improvement in scene classification. Motion based feature vector of the proposed approach captures motion information over a finite space time volume and thus is robust to noise in the given video.

Fig. 1(a,c) show video frames from two different elevator scenes and Fig. 1(e) shows the accuracy of classification obtained using the proposed approach for a number of trials conducted. Another such example is shown in Fig. 1(b,d,f). It can be observed that the result is consistent and high for all the trials. The proposed approach successfully performs the classification task for different scenes even when there is high variability in their appearances.

The primary contribution of the proposed approach are stated below.

- We incorporate both spatial and temporal information in the feature descriptor for efficient scene classification,
- The proposed approach is robust to the quality of video used for developing training model,
- Comparatively, lesser number of training data is required for the scene classification for obtaining good accuracy,
- The main object from the image that defines the semantics of a scene and used for scene classification is assigned the correct class, and
- The proposed approach shows significant improvement in accuracy compared to other methods in scenes with texture and motion.

In section 2, a brief description about the related research work done previously is provided. Section 3 contains a description about the proposed algorithm for classification. Section 4 summarizes the conducted experiment on the proposed approach. In section 5, results of applying the proposed algorithm on a dynamic scene dataset are described. Section 6 discusses some of the challenges and directions for future research. Section 7 provides the conclusion of the present work.

2. Related Work

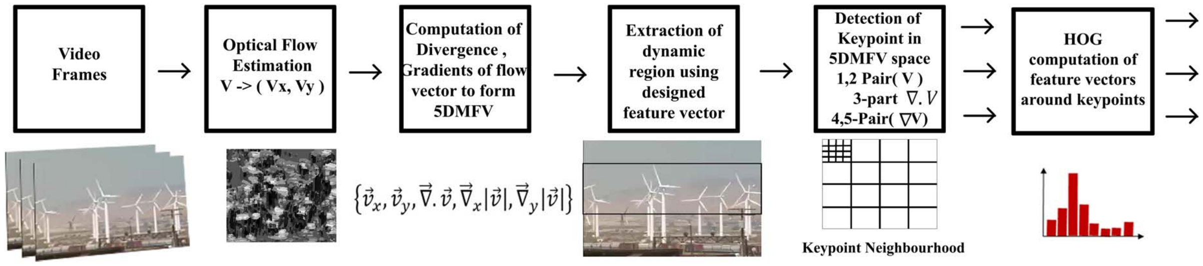
As mentioned in section 1, plenty of spatial scene classification algorithms are available in the literature [10]. The earlier ones include gist based scene classification [22]. In this algorithm, a holistic envelope of an image is used for scene classification. In [5], it is shown that employing histograms of oriented gradients (HOG) results in improved object recognition compared to the other gradient and edge based descriptors such as [21]. In [13], scene classification is dealt with computation of gist globally, also the local textures at various scales. In [26], color descriptor based approach is used to increase illumination constancy. In [11], local spatial features similar to visual codewords are used in constructing models for the scenes. In [16], spatial pyramid is used for identifying the object class. This approach splits the image into fine sub-regions and computes the spatial pyramid by computing histograms for each subregion. In [27], local semantic concepts are used for representing scenes wherever they are present. Thus, all these approaches involve extraction of local spatial descriptors for scene classification task.

In [17], an image is represented as an object bank where it is expressed as scale invariant response function of object detectors. This approach is proved to be better than low level image representation based scene classification and is also proved to be effective in higher level visual recognition tasks. A similar technique is developed for high level representation of activity [23]. Recognition of dynamic textures using histograms of spacetime orientation structure is developed in [7]. In [9], action recognition is developed from inspirations in biological systems. A significant improvement in accuracy on several datasets is reported compared to previous scene classification algorithms.

Some of the spatio-temporal feature based algorithms are used in specific tasks like action recognition. In [8], a spatio-temporal motion descriptor is used for recognising actions. In this algorithm, patterns of motion is recognized from noisy optical flow output and is processed to form the motion descriptor. In [4], spatio-temporal pyramid matching (SPTM) is used for querying video shots. In this method, SPTM is used for improved dynamic object matching, and better results were obtained, compared to the previous methods.

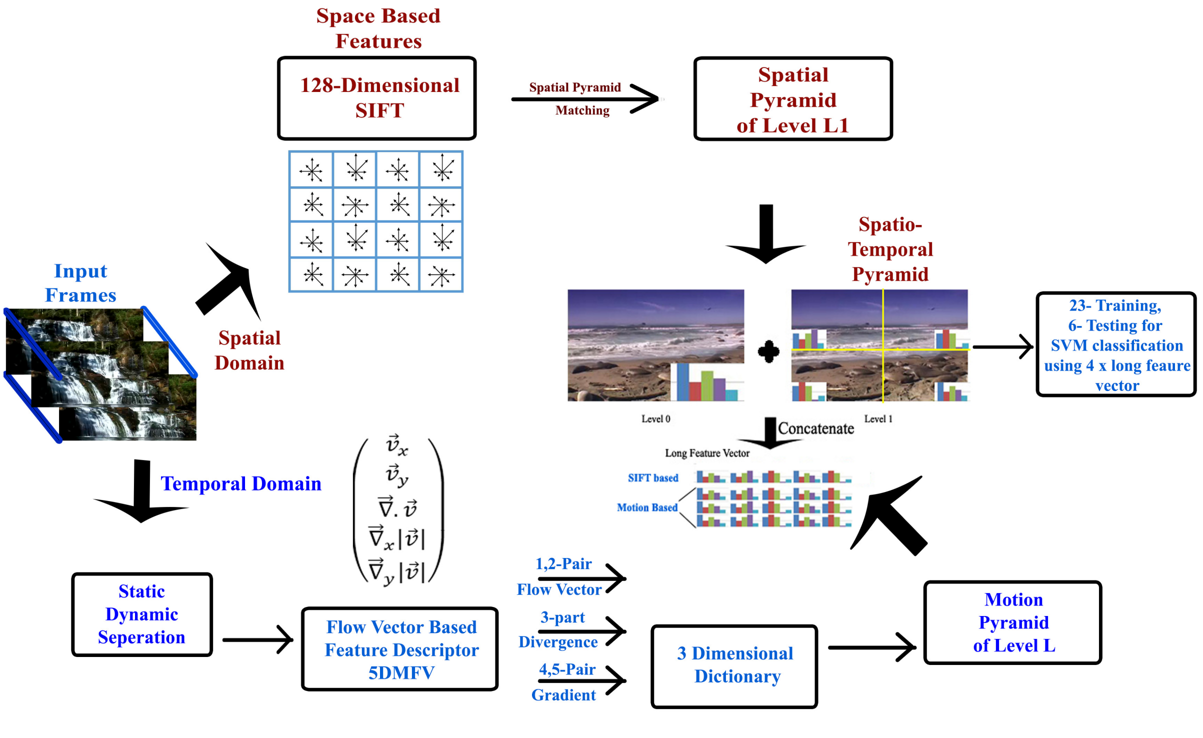
Recent papers have attempted to solve the problem of dynamic natural scene classification. In [6], classification is based on spatio-temporal filters, while [24] extracts dynamic invariants in disordered systems. The other recent work on video scene classification involves a local descriptor based on Slow Feature Analysis (SFA), that represents stable and prime motion components of training video. These local descriptors are then integrated into global coding architecture for providing a model for each semantic category [25].

Lucas-Kanade[19] and Horn-Schunck[15] are the standard optical flow algorithms that act as building blocks to construct the flow vectors of a dynamic scene. These al-


Video Frames
Optical Flow Estimation V -> (Vx, Vy)
Computation of Divergence , Gradients of flow vector to form 5DMFV
Extraction of dynamic region using designed feature vector
Detection of Keypoint in 5DMFV space 1,2 Pair(V) 3-part ∇.V 4,5-Pair(∇V)
HOG computation of feature vectors around keypoints
$\{\vec{v}_x, \vec{v}_y, \vec{\nabla}.\vec{v}, \vec{\nabla}_x|\vec{v}|, \vec{\nabla}_y|\vec{v}|\}$
Keypoint Neighbourhood

(a)

Figure 2: Extraction of 5DMFV from dynamic scene for forming motion descriptor for spatial pyramid matching. Motion descriptor is extracted by applying Histogram of Oriented Gradients (HOG) on 5DMFV.


Space Based Features
128-Dimensional SIFT
Spatial Pyramid Matching
Spatial Pyramid of Level L1
Input Frames
Spatial Domain
Spatio-Temporal Pyramid
23- Training, 6- Testing for SVM classification using 4 x long feaure vector
Level 0
Level 1
Concatenate
Long Feature Vector
SIFT based
Motion Based
Temporal Domain
$\vec{v}_x$ $\vec{v}_y$ $\vec{\nabla}.\vec{v}$ $\vec{\nabla}_x|\vec{v}|$ $\vec{\nabla}_y|\vec{v}|$
Static Dynamic Seperation
Flow Vector Based Feature Descriptor 5DMFV
1,2-Pair Flow Vector
3-part Divergence
4,5-Pair Gradient
3 Dimensional Dictionary
Motion Pyramid of Level L

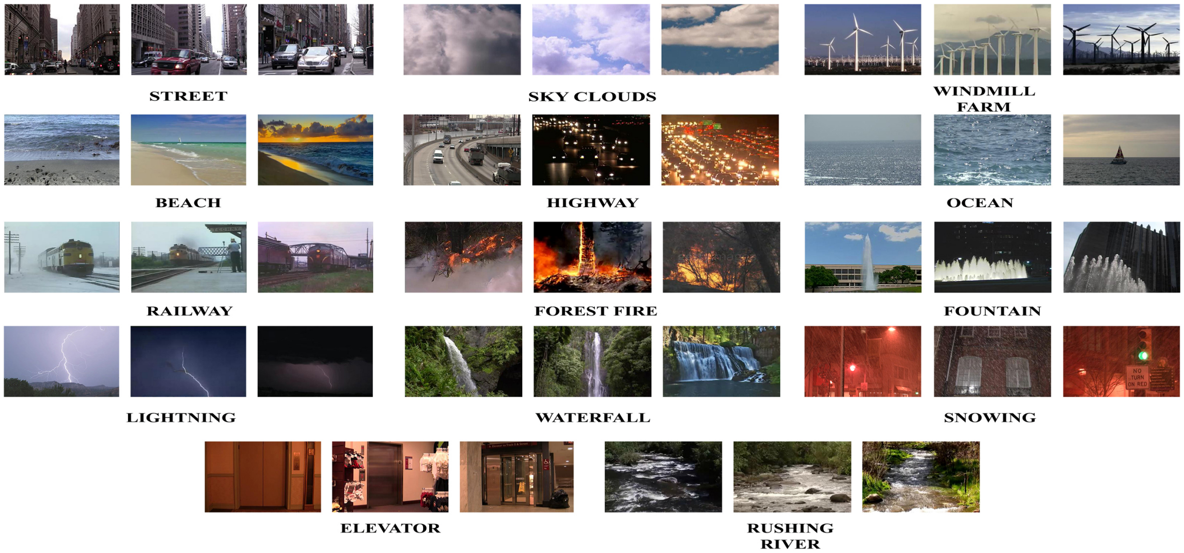
(a)

Figure 3: Illustration of the Proposed Approach.

gorithms 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, high variation in depth, and specular reflections. Some of these shortcomings have been overcome in the approach which uses feature descriptor matching [1]. 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 [2].

3. Proposed Approach

We attempt to exploit both the spatial and temporal information in order to address the problem of scene classi-



(a)

Figure 4: Video categories from UPenn dataset used for classification.

fication. Spatial information is useful in cases where there is high texture gradient in the scene. Motion information is useful when the texture gradient is less and the motion information of the objects in the scene can be estimated. Thus, the information extracted from these two paradigms complement each other and lead to sufficient discrimination between different scene categories.

Fig. 2 depicts the process of construction of temporal feature descriptor and Fig. 3 summarizes the construction of spatio-temporal pyramid that is used by the proposed approach for constructing training models for scene classification. First, a set of frames are extracted at fixed intervals from input video. Flow vectors are estimated between consecutive pair of extracted frames. The motion based feature vector is constructed by computing divergence, gradient and mean of flow vectors across the frames used, and is used to extract the dynamic regions. Finally, histogram of oriented gradients is computed to construct descriptor for motion around the detected keypoints, as depicted in Fig. 2. The computed feature descriptor is used to construct motion pyramid, which is then concatenated with spatial pyramid obtained using SPM algorithm, to form training models for each dataset (Fig. 3). Here, the number of levels in spatial and motion pyramid are denoted by L1, and L respectively.

We begin with the description of SPM and then describe about the spatial and temporal feature descriptors that are used in SPM. We introduce here a novel temporal descriptor for classifying video corresponding to a natural scene.

3.1. Spatial Pyramid Matching

In high level computer vision problems, an image can be described by local features extracted from the patch around a pixel location and edge point local descriptors describe the shape in the image. There are local feature descriptors such as GIST[22] and SIFT[18] descriptors which are subsequently used for the object matching and recognition. These feature descriptors are expressed as a set of vectors in d-dimensional feature space.

$$X = (x_1, x_2, x_3, ..., x_d) \tag{1}$$

Pyramid matching operates in the feature space computing the weighted sum of the number of matches that occur at each scale[12]. At each scale, we have a sequence of coarser grids over the feature space and matching is confirmed if the points are in the same cell of the grid. Though this approach allows the matching of features in a higher dimensional space, it discards the spatial information.

SPM uses a multi resolution histogram pyramid in the feature space to perform a matching between a set of d-dimensional feature vectors. The clustering technique for pyramid matching can be performed in 2D image space and then extended to motion feature space[25]. We extract the local feature vectors in the corresponding feature spaces and quantize them into M types each of which is a code word in the codebook. SPM works in L levels of resolution. In each level l, the image is partitioned into $(2^l)^2$ grids of the same size. For each level of resolution and each channel, SPM counts the intersection of codewords which have fallen into the same grid in the image space to compute the grid similarity of the two images. Finally, we weigh each spatial histogram as[14].

$$K(I_1, I_2) = \sum_{l=1}^{L} \sum_{i=1}^{G(l)} \frac{1}{2^{L-l+1}} K_{l,i}(I_1, I_2) \qquad (2)$$

$$K_{l,i}(I_1, I_2) = \sum_{m=1}^{M} min(H_l^m(I_1), H_l^m(I_2)) \qquad (3)$$

Here, L is the total number of levels and $G(l)$ is the total number of grids in level l. $H_{l,i}(I_1)$ is the number of code word m appearing in the i^{th} grid of the l^{th} level in image I_1. We have chosen $L = 2$ or $L = 3$ while the dictionary size is set to be 200.

3.2. Feature Extraction

This section briefly explains the type of features used for SPM. We use both the spatial and temporal cues from the video frames in order to extract the desired feature descriptor. For spatial information, we use high dimensional strong features which are SIFT descriptors of 16×16 pixel patches computed over a grid with a spacing of 8 pixels. This feature is very similar to the approach of [16]. K-means clustering of a random subset of patches from the training set is performed to construct a visual vocabulary.

For the incorporation of temporal information, a 5-dimensional motion flow vector (5DMFV) from Lucas-Kanade optical flow field [19], its divergence and its gradients are calculated on a set of video frames.

$$P_i(x, y) = (p_1, p_2, p_3, p_4, p_5) \qquad (4)$$

$$Q = (q_1, q_2, q_3, q_4, q_5), \qquad (5)$$

where $Q_j = \log(\sigma^2(P_{ji})), j = 1, 2, 3, 4, 5.$

In equation (4), (p_1, p_2) = $(\vec{v}_x, \vec{v}_y)$ are the optical flow vectors obtained from Lucas-Kanade algorithm. p_3 corresponds to the divergence $(\dot{\bigtriangledown}\vec{v})$ while (p_4, p_5) are the gradients of the magnitude of the vector field $\vec{v}$ in x and y directions.

Equation (5) represents 5DMFV where σ^2 is the variance of the features on a finite time support. This 5DMFV helps in the segmentation of video frames into static and dynamic regions. We use variances of flow vector statistics for the construction of feature vector so that these are consistently zero in static regions, regardless of the amount of noise present in the input video. The intuition behind using the logarithm is to capture subtle motions present in the scene, so that we are successfully able to detect and extract important dynamic object using the feature vector. These intuitions are further validated in the results section.

As the static region provides no useful information of the motion characteristics of the scene, we are interested only in the temporal properties of the dynamic regions. For this purpose, we use the 5DMFV to separate static and dynamic regions and then extract 5DMFV for the separated dynamic region using a bounding box. For each 5DMFV derived from the set of video frames, we derive feature descriptor on 16×16 patches over a grid with spacing of 1 pixel on the motion vector space. In each of the patches, we take the gradients on the resultant of q_1 and q_2 $(\vec{v})$, q_4 and q_5 $(\bigtriangledown\vec{v})$, and q_3 on 5DMFV as (q_1, q_2) and (q_4, q_5) are x and y components: Gradient directions and magnitude is taken around each keypoint which is the centre of the patch. In this histogram, 360 degrees of orientation are broken into 8 bins. A histogram is created with 8 bins with each storing an amount proportional to the magnitude of gradient at that point. This feature descriptor looks similar to the SIFT descriptor except for the fact that we use 5DMFV space instead of the 2D image space and we avoid difference of Gaussian on 5DMFV as it is less suitable in this space.

The size of the pyramid depends on the dictionary size and the number of pyramid levels. Let us assume that there are l levels. For the i^{th} level of pyramid, number of blocks is given by:

$$N(i) = (2^i)^2 \qquad (6)$$

If a dictionary level of size D is used for each level, then total size of pyramid, L is given by:

$$L = D \times \sum_{i=0}^{l} (2^i)^2 \qquad (7)$$

Thus, if we have number of levels, l =3, and a dictionary of size 200, size of pyramid is given by:

$$L = 200 \times (16 + 4 + 1) = 4200. \qquad (8)$$

For a patch size of 16, distance between grid centres of 1, image size of 256×256 and number of visual words of 200, 4200- dimensional feature descriptor is obtained for each vector in 5DMFV. Finally, we have feature descriptor of one 4200- dimensional from spatial and three 4200-dimensional vectors for each vector. In a simple way, we can either concatenate spatial and temporal domain feature descriptor to make a 13800 dimensional vector or a 4200×4 matrix.

4. Experiment

We used the UPenn dataset[1] with 14 different categories of video scenes having 30 videos in each of these categories. Each video has about 150 frames running for 5 seconds depicting a unique natural scene. Fig. 4 shows a set of three

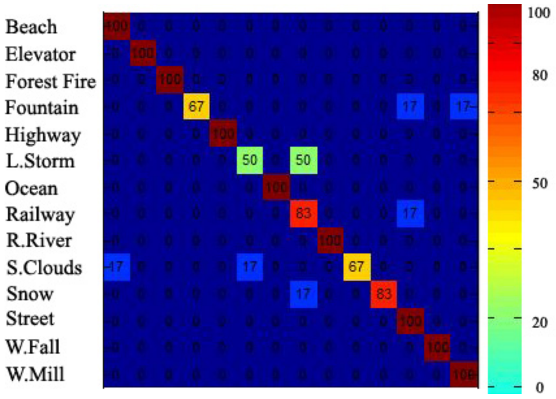
[1] http://www.cse.yorku.ca/vision/research/dynamic-scenes.shtml

Method	Average Accuracy	Highest Accuracy
Resultant of flow vector	69.81± 6	79.76
Divergence	66.21± 3.03	73.81
Resultant of gradient	59.045± 4.5	66.67
5DMFV	76.44± 4.43	84.52
SIFT	80.18± 3.18	88.1
SIFT+5DMFV	84.27± 2.94	90.48

Table 1: Column-2 provides average accuracy along with standard deviation obtained in multi class classification, column-3 gives maximum accuracy in each category.

representative images from all of the datasets that are used for testing the proposed approach and are part of the UPenn dynamic scenes dataset [25]. We extract a frame from each video for applying spatial pyramid matching on the SIFT descriptor. In the case of temporal space, we derive the 5DMFV from a set of frames for each video. Feature descriptor of 4×4200 matrix is computed as mentioned in the proposed approach. Having obtained feature descriptor for every video in all categories, multi class classification is done using Support Vector Machine (SVM)[3]. We have noted that the average of per class recognition along with standard deviation of the results from the individual runs. We represent this classification in confusion matrix as shown in Fig. 5 where the diagonal element values depict the accuracy of classification of a particular category in the testing set of videos.

5. Results



(a)

Figure 5: Confusion matrix for the UPenn dataset obtained with average accuracy with 89.29

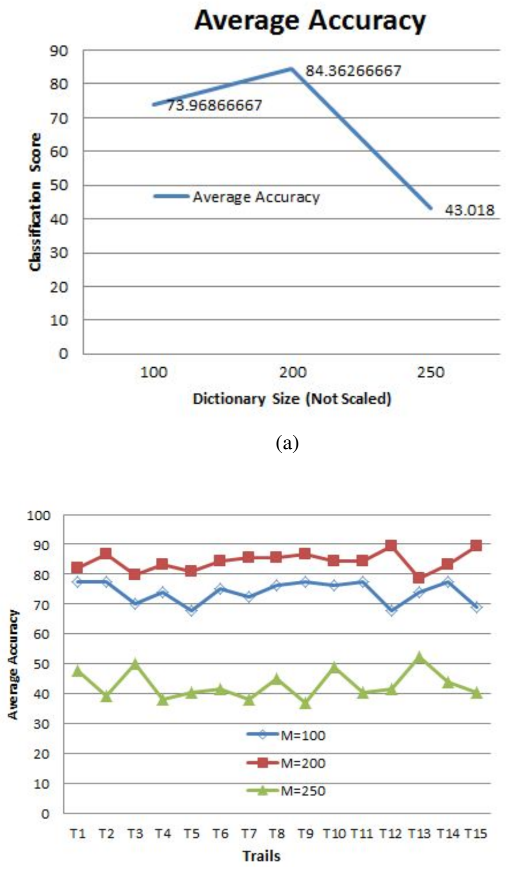
Scenes	HOG [20, 6]	GIST [17, 6]	Chaos [24]	SOE [6]	SFA [25]	Our Model
Beach	37	90	27	87	96	98.3
Elevator	83	50	40	67	86	89.8
F.Fire	93	53	50	83	90	80
Fountain	67	50	7	47	63	60.1
Highway	30	40	17	77	70	88.1
L.Storm	33	47	37	90	80	66.6
Ocean	47	57	43	100	96	89.8
Railway	60	93	3	87	83	86.5
R.River	83	50	3	93	83	95
S.Clouds	37	63	33	90	100	91.5
Snow	83	90	17	33	73	89.9
Street	57	20	17	83	90	96.6
W.Fall	60	33	10	43	86	74.9
W.Mill	53	47	17	57	90	91.5
Average	59	56	20	74	85	85.61

Table 2: Classification results in average accuracy for the UPenn dataset

Table 1 shows the results of classification experiments using 23 videos for training and 6 videos for testing in each class. Number of levels, $L = 2$ and dictionary size, $M = 200$. We use $L = 2$, as single level performance drops as we go from $L = 2$ to $L = 3$ in the usage of strong features, adding to the disadvantage of excess time complexity and memory. Visual vocabulary size, $M = 200$ is assigned for the above experiment as $M = 400$ yields little improvement in the performance. We have listed the performance of different feature descriptors such as feature descriptor from flow vector, its divergence, its gradients, concatenated flow vectors with its divergence and gradients, SIFT on image space. We can see the significant improvement in the accuracies as we introduce both spatial and temporal data. Confusion matrix for the UPenn dataset is shown in Fig. 5. From the confusion matrix, we observe that except for a scene like lightning, where there is a large variation in brightness level, the classification accuracy is high for all the scenes.

Table 2 reports the comparison of our model with the state-of-the-art results: HOG[20, 6], GIST[17, 6], Chaos[24], SOE[6], and Slow Feature Analysis(SFA)[25]. Though there are variations in accuracies in individual categories, our model's average final accuracy appears similar to [25] model. Experiment is repeated ten times randomly choosing the training set for computing the average accuracy for each category. Our model yield an overall average accuracy of 85.61.

We observe from the graph Fig. 6(a) that we obtain the maximum average accuracy for scene classification for dictionary size, M=200 and accuracy deteriorates for increase or decrease in M. Thus, we can conclude that coarse-grained spatio-temporal cues have more discriminative powers than enlarged vocabulary. Graph in Fig. 6(b) shows that the average accuracy results from the experiment



(a)

(b)

Figure 6: (a) Graph depicts the variation of average accuracy in multi class classification for the variation of dictionary size ($M = 100, 200, 250$), (b) Here we observe the consistency in the experiment for different dictionary sizes.

remains consistent for fifteen trails, for a fixed dictionary size. Though we consider random training sets from each video category for each trial, results were found to be consistent.

6. Challenges

Motion feature descriptor is derived from optical flow vector which shows less accurate results in changing illumination in the scene . The proposed approach is not optimized for the value of dictionary size that yields maximum accuracy. Also, computation of temporal feature vector is time consuming. Future work could be investigating the 5DMFV for the high variation in sampling in video for frame extraction, and the elimination of the current limitations of the approach.

7. Conclusion

The combined feature descriptor from spatial and temporal domain for spatial pyramid matching yields a good multi class SVM classification. SIFT feature descriptor is used for spatial domain while 5DMFV is extracted for dynamic motion information. Even in cases where lighting condition changes significantly, the proposed approach is able to produce result. Thus, the approach shows comparable results with the state-of-the-art methods for the scene classification from videos. The proposed approach for scene classification from natural videos has a variety of applications such as video retrieval, video tagging, and action recognition. We hope to improve the accuracy of the proposed approach in the future and employ it to solve other related applications in computer vision.

References

[1] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer vision and image understanding*, 63(1):75–104, 1996. 3

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

[3] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. 6

[4] J. Choi, W. J. Jeon, and S.-C. Lee. Spatio-temporal pyramid matching for sports videos. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, MIR '08, pages 291–297, New York, NY, USA, 2008. ACM. 2

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 2

[6] K. G. Derpanis, M. Lecce, K. Daniilidis, and R. P. Wildes. Dynamic scene understanding: The role of orientation features in space and time in scene classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1306–1313. IEEE, 2012. 2, 6

[7] K. G. Derpanis and R. P. Wildes. Dynamic texture recognition based on distributions of spacetime oriented structure. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 191–198. IEEE, 2010. 2

[8] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 726–733. IEEE, 2003. 2

[9] M.-J. Escobar and P. Kornprobst. Action recognition via bio-inspired features: The richness of center–surround interaction. *Computer Vision and Image Understanding*, 116(5):593–605, 2012. 2

[10] L. Fei-Fei, R. Fergus, and A. Torralba. Recognizing and learning object categories. *CVPR Short Course*, 2007. 2

[11] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. IEEE, 2005. 2

[12] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1458–1465. IEEE, 2005. 4

[13] S. Grossberg and T.-R. Huang. Artscene: A neural system for natural scene classification. *Journal of Vision*, 9(4), 2009. 2

[14] J. He, S.-F. Chang, and L. Xie. Fast kernel learning for spatial pyramid matching. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7. IEEE, 2008. 5

[15] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1):185–203, 1981. 2

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006. 2, 5

[17] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, pages 1378–1386, 2010. 1, 2, 6

[18] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 4

[19] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 81, pages 674–679, 1981. 2, 5

[20] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2929–2936. IEEE, 2009. 6

[21] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(4):349–361, 2001. 1, 2

[22] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001. 1, 2, 4

[23] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1234–1241. IEEE, 2012. 1, 2

[24] N. Shroff, P. Turaga, and R. Chellappa. Moving vistas: Exploiting motion for describing scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1911–1918. IEEE, 2010. 2, 6

[25] C. Theriault, N. Thome, and M. Cord. Dynamic scene classification: Learning motion descriptors with slow features analysis. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2603–2610. IEEE, 2013. 2, 4, 6

[26] K. E. Van De Sande, T. Gevers, and C. G. Snoek. Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1582–1596, 2010. 2

[27] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, 72(2):133–157, 2007. 2