SA-CNN: Dynamic Scene Classification using Convolutional Neural Networks

Aalok Gangopadhyay, Shivam Mani Tripathi, Ishan Jindal, Shanmuganathan Raman

*Indian Institute of Technology Gandhinagar, Gujarat, India*

**Abstract**

The task of classifying videos of natural dynamic scenes into appropriate classes has gained lot of attention in recent years. The problem especially becomes challenging when the camera used to capture the video is dynamic. In this paper, we propose a statistical aggregation (SA) solution based on convolutional neural networks (CNNs) to address this problem. We call our approach as SA-CNN. The algorithm works by extracting CNN activation features for a number of frames in a video and then uses a statistical aggregation scheme in order to obtain a robust feature descriptor for the video. We show through results that the proposed approach performs better than the state-of-the-art algorithm for the Maryland dataset. The final descriptor obtained is powerful enough to distinguish among dynamic scenes and is even capable of addressing the scenario where the camera motion is dominant and the scene dynamics are complex. Further, this paper shows an extensive study on the performance of various statistical aggregation methods and their combinations in order to obtain minimal classification error. We compare the proposed approach with other dynamic scene classification algorithms on two publicly available datasets - Maryland and YU Penn to demonstrate the superior performance of the proposed approach.

**Keywords:** Dynamic Scene Understanding, Video Classification, Convolutional Neural Networks, Deep Learning

1. **Introduction**

Consider the video of a natural dynamic scene. The video could have been captured either by a static or a dynamic camera. Given several categories comprising of natural scene videos, we would like to assign the correct category for a given video. This dynamic scene classification problem is more challenging for a moving camera than that for a static camera.
In the case of images, a lot of significant research has been done to address the problem of scene recognition. Image scene recognition involves classifying an image into one of the several given classes (SUN Database) \[1\]. Convolution Neural Network (CNN) based approaches have recently dominated the task of image scene classification, obtaining very high accuracy and outperforming other previous state-of-the-art approaches by a significant margin. These approaches have worked remarkably well on several other large scale image datasets with up to thousands of categories. These powerful methods focus on finding appropriate spatial feature descriptors for a given image. Hence, they take into consideration only the spatial description of the scene present in the image.

In contrast to image scene classification where the class labels are based only on the spatial properties, dynamic scene classification tries to classify videos into different categories whose semantic labels is derived from the activities occurring in the scene. Several examples of dynamic scenes are shown in Figure 1. The dynamic scenes like 'avalanche' is given its class label based on the movement of ice, and not just based on the spatial attributes of the scene.

![Figure 1: Example frames of classes within (a) YUPenn dataset (left) and (b) Maryland dataset (right).](image)

The proposed approach is inspired by the unparalleled success of Convolutional Neural Network (CNN) based approaches for various recognition tasks over the past few years. Krizhevsky et al. mentioned the idea of using very large and deep CNN models to classify videos as well \[2\]. But building new architecture for videos and training it on a very large dataset is a complex procedure. However, two recent works on large scale video classification use CNNs to achieve the task by learning features from hundreds of thousands of videos extracted from Youtube or Facebook \[3\] \[4\]. In \[3\], the proposed architecture was trained on a large collection of sports videos (Sports-1M) for about a month and obtained very good results when tested on UCF-101.

A number of CNN implementations pretrained on a large database of images are available which are ready to be used for off-the-shelf image feature extraction. We take forward this idea of classification with off-the-shelf CNN implementation in order to perform dynamic scene classification for videos by
using CNN model Caffe (AlexNet) [5] on frames of videos. The present approach differs from that of [3] in the sense that the CNN trained on standard image dataset (ImageNet) is used to classify videos of dynamic scenes. This relieves us from training CNN on a new video dataset. We use very simple yet effective tools for dynamic scene classification and show that even common statistical measures can be employed to capture the temporal variation which can be combined with spatial information for dynamic scene videos. We enhance this framework and adapt it for the problem of dynamic scene classification and obtain very high accuracy.

It is worth mentioning that all of the previous dynamic scene classification methods, except Tran et al.’s C3D [4], relied on using local features and did not exploit very large dataset. The proposed approach uses CNN which has been pre-trained on ILSVRC 2012 dataset, which has around 1 million images spanning across 10,000 categories. However, the database is largely dedicated to object recognition tasks but not dynamic scene classification tasks [6]. The primary contribution of the proposed approach are listed below.

1. Exploiting pre-trained CNN models and adapting it to the dynamic scene classification task for obtaining maximum accuracy,
2. Using common statistical measures to merge spatial features with their temporal variation to arrive at a novel feature descriptor,
3. We design the classification algorithm in such a way that it is highly robust to scene motion as well as camera motion,
4. We obtain state-of-the-art result on two dynamic scene datasets - Maryland and YUPenn. The increase in classification accuracy is observed to be very high in the case of Maryland dataset.

The rest of the paper is organized as follows. Section 2 describes the relevant works about dynamic scene classification and the CNN literature in detail. Section 3 presents a detailed account of the proposed approach. In section 4, we present results and comparisons of the experiments carried out with complete quantitative analysis. We conclude the paper in section 5 with some suggestions for future work.

2. Related Work

A lot of work has been done in the field of scene classification in the past decade. This includes recognizing scenes from single images as well as classifying dynamic scenes from videos. In this section, we shall elaborate on some of the past works which are directly related to the present work.

In the field of single image recognition tasks, bag-of-features based methods were initially prevalent among the research community [7, 8, 9, 10, 11]. These methods were based on the principle of geometric independence or orderless spatial representation. Later, these methods were enhanced by the inclusion of weak geometric information through spatial pyramid matching (SPM) [12]. This method employed histogram intersection at various levels of the pyramids for matching features. However, CNN based approaches have been able to achieve even higher accuracies as observed in some of the recent works [13, 4, 14, 15].
This sparked a lot of recent research work on architectures and applications of CNNs for visual classification and recognition tasks.

In [2], the CNN architecture consisted of five convolutional layers, followed by two fully connected layers (4096 dimensional) and an output layer. The output from the fifth max-pooling layer was shown to still preserve global spatial information [15]. Even the activation features from the fully connected layer were found to be sensitive to global distortions such as rotation, translation, and scaling [13]. However, they have proven to be very powerful general feature descriptors for high level vision tasks. Several CNN implementations such as DeCAF, Caffe and OverFeat, trained on very large datasets are available for feature extraction to perform image classification tasks [16, 5, 14]. These CNNs, pre-trained on large datasets such as ImageNet, have been efficiently used in scene classification and have achieved high/impressive accuracies [13] (for example, MOP CNN, OverFeat, etc.). Also, the ImageNet trained model of these implementations have been shown to generalize well to accommodate other datasets as well [16, 15]. CNN features obtained from object recognition datasets have also been used for obtaining high accuracy in various high level computer vision tasks [17].

On the other hand, research in dynamic scene classification from videos has been dominated by the idea of finding more powerful and robust local spatio-temporal feature descriptors. This is followed by embedding weak global information to find most appropriate representation of the given video. Initially, spatial and temporal feature based approaches such as GIST+HOF and GIST+Chaos were employed to perform dynamic scene classification [18, 19, 20]. In [18], it was shown that spatial and temporal descriptors together gave better results than using either of them alone. These methods were built and tested for Maryland (In-the-Wild) dataset introduced by [19].

The spatio-temporal based approaches were introduced by spatio-temporal oriented energies [21], which also introduced the YUPenn dataset. The very same work concluded that even relatively simple spatio-temporal feature descriptors were able to achieve consistently higher performance on both YUPenn and Maryland datasets as compared to HOF+GIST and HOF+Chaos approaches. More details for both the dynamic scene datasets have been covered in Section 4.1. Current state-of-the-art approach, bags of space-time energies (BoSE), proposes using a bag of visual words for dynamic scene classification [22]. Here, local features extracted via spatio-temporally oriented filters are employed. They are encoded using a learned dictionary and then dynamically pooled. The technique currently holds the highest accuracy on the two mentioned datasets [22] amongst all peer-reviewed studies. Recently, a work done by Duran et al. (not peer-reviewed yet) uses a novel three dimensional CNN architecture for spatio-temporal classification problems [4]. This technique shows promising results and marginal improvement over current state of art method.
3. Proposed Approach

Most of the recent works on dynamic scene classification have focused on dense extraction of spatio-temporal features, followed by feature encoding and pooling strategies to obtain the final feature representation for a video. Several other methods have considered separately extracting spatial and temporal features and then combining them to obtain the final feature representation for a given video. However, we use a different approach here. Given a video, we first extract spatial feature descriptors for a chosen number of frames. After that, we use aggregation strategies to obtain information about the temporal variation of the spatial features. Using this information, we arrive at the final feature descriptor for a given video to be classified. The entire process has been outlined in Figure 2. In this section, we shall describe the proposed approach in more detail.

3.1. Feature Extraction from Frames

To start with, we extract feature descriptors for the frames of a video using spatial information only. Given a video $V_k$ containing a total of $N_{0k}$ frames, we select $N_k$ frames that are linearly spaced in the interval $[1, N_{0k}]$. An important thing to note here is that the temporal distance between consecutive frames in the set of $N_k$ frames differ from video to video, as there is a lot of variation in the frame rate and the total video duration in the datasets. After selection, each of these $N_k$ frames are resized to a resolution of $256 \times 256$. Then we
extract the CNN activation features for each of these frames using CNN Caffe implementation - AlexNet \cite{5}, which has been trained on ILSVRC 2012 dataset of images \cite{6}. Please note that this model has minor variations from the model mentioned in \cite{2}. For each of the frame $f_i$ taken as input, we take the output of the seventh fully connected layer of the CNN (post ReLU transformation) and obtain the 4096 dimensional feature descriptor $X_i$. Thus, after the feature extraction step, we obtain an $N_k \times 4096$ dimensional matrix $X$ for the video $V_k$. This matrix contains the information about how each of these 4096 features evolve with time and hence can be exploited to extract the temporal properties. We have 4096 time curves each of length $N_k$. Ideally we would like to use a feature descriptor for each of these curves that captures the temporal variations in a robust way. But such an ordered temporal descriptor would make the final descriptor for the video very huge, since there are 4096 such curves. Hence, rather than using the ordered properties, we use temporally orderless statistics for each curve. We show that simply using the first few moments yield very high accuracy for both the datasets.

3.2. Aggregation

From the previous step, we obtain a set of 4096 dimensional vectors, $\{X_1, ..., X_K\}$, each of which contains rich spatial information and represents a single frame in the given video. In this step, we combine these $K$ vectors $\{X_1, ..., X_K\}$ in time in order to capture the temporal statistics of the spatial features. We do this to induce a certain degree of temporal invariance and extract temporally orderless properties. For this, we consider two strategies to aggregate these spatial descriptors temporally.

1. Statistical Aggregation (SA): The simplest method is to use statistical measures like moments, for $M = N_k$ instances for each of the 4096 dimensions. Let $X_{ij}$ denote the $i^{th}$ instance of $j^{th}$ dimension of the feature descriptor, where $i \in \{1, 2, ..., M\}$ and $j \in \{1, 2, ..., 4096\}$. Then we use the following statistical measures to aggregate $M$ instances to get the final feature descriptors.

   (a) Mean: $\hat{\mu} = (\mu_1, \mu_2, ..., \mu_{4096})$ where: $\mu_j = \frac{1}{M} \sum_{i=1}^{M} X_{ij}$

   (b) Standard Deviation: $\hat{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_{4096})$ where: $\sigma_j = \sqrt{\frac{1}{M} \sum_{i=1}^{M} (X_{ij} - \mu_j)^2}$

   (c) Skewness: $\hat{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_{4096})$ where: $\gamma_j = \frac{E[(X_j - \mu_j)^3]}{\sigma_j^3}$

   (d) Kurtosis: $\hat{\kappa} = (\kappa_1, \kappa_2, ..., \kappa_{4096})$ where: $\kappa_j = \frac{E[(X_j - \mu_j)^4]}{\sigma_j^4}$

Please note that each of $\hat{\mu}, \hat{\sigma}, \hat{\gamma}$ and $\hat{\kappa}$ are 4096 dimensional vectors. For classification, we can consider each one of these moments individually or their various combinations (concatenation) to get the final feature descriptor of the video. Details of the combinations used have been covered in section \cite{4}.
2. **Robust PCA + Statistical Aggregation:** The $4096 \times M$ dimensional representative feature descriptor can be considered as a data matrix which comprises of the feature vector of each frame of the video as columns. Thus we obtain a data matrix $D$ which in most cases has full rank as $M << 4096$. This matrix can be considered to be corrupted by noise arising due to temporal dynamics, camera jitter/motion, etc. It is difficult to understand in what sense this matrix is depicting the spatio-temporal information. To get the latent information from the matrix, we break the matrix using robust principal component analysis (RPCA) into sum of two matrices [23]. Taking inspiration from the effective results of Robust PCA in background subtraction [24] which works on pixels, we used augmented Lagrangian multiplier technique [25] to decompose the data matrix of CNN features $D$ into a rank deficient matrix $\hat{A}$ that describes the underlying scene spatially and a sparse error matrix $\hat{H}$ which contains temporal information in some form as seen in equation (1).

$$D = \hat{A} + \hat{H}$$  \hspace{1cm} (1)

The aggregation method proposed in the previous subsection works well when directly applied to the data matrix $D$. Also, similar results are obtained when applied to the rank deficient matrix $\hat{A}$. But our aim in dynamic scene classification is to make use of the temporal information in order to improve the classification results. So, we try to use this full rank sparse matrix $\hat{H}$ as the representative of temporal information. We incorporate this information in the classification task by column wise concatenation of the matrix $\hat{A}$ and $\hat{H}$. In this manner, we obtain a feature descriptor describing the dynamic scene and perform statistical aggregation again as mentioned above.

In the next section, we shall explain the various experiments we carried out and show their results with both Direct Statistical Aggregation and robust PCA + Statistical Aggregation. We also show the comparison with other competing methods to emphasize the significance of the contribution in this work.

4. **Results and Discussion**

4.1. **Dataset**

We evaluate our method on the following two datasets.

1. **YUPenn** (Stabilized Dataset) : This dataset contains 14 classes with 30 videos in each class making it a total of 420 videos. Each video contains around 145 frames on an average, with the frame rate not being the same for each video [21].

2. **Maryland** (In-the-Wild Dataset) : This dataset contains 13 classes with 10 videos in each class making it a total of 130 videos. Each video has the same frame rate of 30 fps. But the duration of the videos and hence the total number of frames varies a lot in the dataset [18].
Both the datasets have large variation in illumination, image scale, camera viewpoint, frame resolution, duration of the video, etc. The videos in Maryland dataset contain large camera motion and scene cuts, whereas those in YUPenn involve static camera and hence contain no camera motion. All the above factors result in large intraclass variations which make dynamic scene classification a challenging task.

4.2. Results

As mentioned in the proposed approach, given a video $V_k$, we extract the 4096 dimensional activation features for each of the $N_k$ frames chosen from the video. Thus we obtain $N_k$ such 4096 dimensional feature vectors for the video $V_k$. To find the final vector representation for each video, we use simple temporal aggregation techniques using statistical moments. After obtaining the feature vector representation for each video in the dataset we perform multi-class classification. For classification, we use one-vs-rest SVM with the leave-one-video-out (LOVO) method as done in previous works [22]. For the implementation of SVM, we use the LIBSVM library[26]. It is found that in the case of Maryland dataset, linear kernel gives best results and in the case of YUPenn, histogram intersection kernel(HIK) works the best.

To analyse the performance of the statistical moments, we initially utilize all the frames, that is, we set : $N_k = N_0k$. We later show that we get good results even for smaller values of $N_k$. Table 1 depicts the results obtained on using statistical moments and their various possible combinations.

It is very surprising to see that apart from mean, even the other statistical moments, such as S.D., skew and kurtosis perform very well when considered individually (as seen in block (a) of table 1). This means that the temporal statistics of the 4096 dimensional vector is highly similar for videos of the same class. This indicates that various types of probability based approaches can be explored for obtaining a very powerful descriptor for the videos. Quick observation of the table reveals that concatenation of all the four moments: mean, standard deviation (S.D.), skew and kurtosis consistently gives high scores on both the datasets. Also, all the feature descriptors which contain either of mean, S.D or their concatenation perform well.

In Table 2 we do a very similar analysis for Robust PCA + aggregation scheme as we did in Table 1 for matrices $\hat{A}$, $\hat{E}$ and their concatenation $\hat{A}$ concat $\hat{E}$. It can be observed that $\hat{A}$ concat $\hat{E}$ and $\hat{A}$ give consistently high scores for most of the statistical aggregation schemes and adding extra moments creates marginal or no difference on the scores. Also, the best score of RPCA aggregation method for any statistical scheme is lesser than or equal to the simple statistical aggregation using first four moments. Moreover, computing RPCA is an expensive procedure than statistical aggregation. Thus we use statistical aggregation scheme ($\hat{\mu} + \hat{\sigma} + \hat{\gamma} + \hat{\kappa}$) instead of RPCA method for its simplicity as well as computational efficiency.

Since the time taken for computing the final feature vector for a given video $V_k$ largely depends on the value of $N_k$, it is very important to understand how the accuracy of the classifier varies with $N_k$. Moreover, to make the computation
| Statistical Measures | Dim  | Yupenn | Maryland |
|----------------------|------|--------|----------|
| Mean                 | 4,096| 96.90  | 90.00    |
| S.D.                 | 4,096| 97.14  | 88.46    |
| Skew                 | 4,096| 88.33  | 74.61    |
| Kurtosis             | 4,096| 89.52  | 63.30    |
| Mean+S.D.            | 8,192| 97.61  | 90.00    |
| Mean+Skew            | 8,192| 96.67  | 91.53    |
| Mean+Kurtosis        | 8,192| 96.67  | 90.76    |
| S.D.+Skew            | 8,192| 96.67  | 90.76    |
| S.D.+Kurtosis        | 8,192| 96.90  | 90.76    |
| Skew+Kurtosis        | 8,192| 89.52  | 73.01    |
| Mean+S.D.+Skew       | 12,288| 97.38  | 90.76    |
| Mean+S.D.+Kurtosis   | 12,288| 97.38  | 90.76    |
| Mean+Skew+Kurtosis   | 12,288| 95.95  | 90.00    |
| S.D.+Skew+Kurtosis   | 12,288| 96.42  | 89.23    |
| Mean+S.D.+Skew+Kurtosis | 16,384| **97.14** | **91.53** |

Table 1: Accuracy obtained for the various moments and their combinations: The first block (a) shows the results of using the first four moments individually. Each of these vectors have a dimension of 4096. In the second block (b), doublet combinations of the moments are obtained by concatenation. Each of these vectors have a dimension of 8192. The third block (c), triplet combinations are considered, each of them having a dimension of 12288. The fourth block (d) shows the result obtained on combining all the four moments resulting in a vector of dimension 16384.
Table 2: Comparison of different methods of classification using RPCA

| Aggregation          | YUPenn | Maryland | YUPenn | Maryland | YUPenn | Maryland |
|----------------------|--------|----------|--------|----------|--------|----------|
| Mean                 | 96.67  | 90.77    | 96.67  | 89.23    | 95.48  | 85.38    |
| S.D.                 | 97.14  | 90.77    | 95.48  | 89.23    | 96.67  | 87.69    |
| Skew                 | 86.67  | 70.77    | 69.29  | 65.38    | 90.95  | 73.08    |
| Kurtosis             | 96.67  | 90.77    | 96.67  | 89.23    | 95.48  | 85.38    |
| Mean + S.D.          | 97.38  | 90.77    | 96.67  | 90.77    | 96.19  | 87.69    |
| Mean + Skew          | 96.90  | 90.00    | 96.19  | 89.23    | 95.24  | 83.85    |
| Mean + Kurtosis      | 96.67  | 90.00    | 96.67  | 89.23    | 95.48  | 83.85    |
| S.D. + Skew          | 97.14  | 90.00    | 94.52  | 89.23    | 95.71  | 87.69    |
| S.D. + Kurtosis      | 97.38  | 90.77    | 96.67  | 90.77    | 96.19  | 87.69    |
| Skew + Kurtosis      | 96.90  | 90.00    | 96.19  | 89.23    | 95.24  | 83.85    |
| Mean + S.D. + Skew   | 97.14  | 90.77    | 96.67  | 90.77    | 95.95  | 86.92    |
| Mean + S.D. + Kurtosis| 97.14 | 90.77    | 96.67  | 90.77    | 95.71  | 84.62    |
| Mean + S.D. + Skew + Kurtosis | 97.38 | 90.00 | 96.90 | 90.77 | 95.43 | 89.23 |
| Mean + S.D. + Skew + Kurtosis | 97.14 | 90.77 | 96.67 | 90.77 | 95.95 | 86.92 |

Table 2: Comparison of different methods of classification using RPCA

time independent of the duration or the frame rate of the video, it would be better to choose the same value of $N_k$ for all videos. Hence, we take same $N_k = N$ for all videos. For obtaining the relation between the accuracy and the number of frames selected from each video, we perform a multiple trials based experiment. In this simple experiment, we evaluate our method for the following values of $N \in \{1, 2, 3, 5, 10, 15, 20, 30, 40, 50, 60\}$. For each of these values of $N$, one-vs-rest SVM with leave-one-video-out (LOVO) is performed 18 times. In each trial $N$ frames are randomly chosen from the video and the 4096 dimensional feature vectors are aggregated using their temporal mean. We evaluate this behavior only for mean, since, for evaluating the accuracy in case of higher moments we need substantial number of frames and hence can’t test them for very low number of frames like 1, 2, 3 and 5.

The graph in figure 3 shows the accuracies obtained in each trial for $N \in \{1, 2, 3, 5, 20, 60\}$. As the value of $N$ increases, the variation in accuracy across the trials decreases. It is apparent from the graph that, for smaller values of $N$, the accuracy is very sensitive to which frame is randomly chosen. Thus such a high fluctuation in accuracy is observed.

The graph in figure 4 depicts how the average accuracy obtained for mean aggregation varies with the number of frames. As expected the accuracy increases as the number of frames increase and saturates around 90% for Maryland and around 97% for YUPenn. Also, it shows that the confidence in accuracy in-
| Class          | HOF + Chaos + GIST | SFA | CSO | BoSE | C3D | SA-CNN [\mu, \sigma, \gamma, \kappa] |
|---------------|---------------------|-----|-----|------|-----|---------------------------------|
| Avalanche     | 20 60 40 60 60 60   | NA  | 90  |
| Boiling Water | 50 60 50 70 80 70   | NA  | 90  |
| Chaotic Traffic| 30 70 60 80 90 90   | NA  | 100 |
| Forest Fire   | 50 60 10 10 80 90   | NA  | 90  |
| Fountain      | 20 60 50 50 80 70   | NA  | 90  |
| Iceberg Collapse | 20 50 40 60 60 60   | NA  | 100 |
| Landslide     | 20 30 20 60 30 60   | NA  | 80  |
| Smooth Traffic| 30 50 30 50 50 50   | NA  | 80  |
| Tornado       | 40 80 70 70 80 90   | NA  | 90  |
| Volcanic Eruption | 20 70 10 80 70 80   | NA  | 100 |
| Waterfall     | 20 40 60 50 50 100  | NA  | 90  |
| Waves         | 10 80 50 60 80 90   | NA  | 100 |
| Whirlpool     | 30 50 70 80 70 80   | NA  | 90  |
| Overall       | 33.08 58.46 43.08   | 60.00 | 57.69 | 77.69 | 91.53 |

Table 3: Comparison of classification scores for various methods of Dynamic Scene Recognition on Maryland dataset.

| Class          | HOF + Chaos + GIST | SFA | CSO | BoSE | C3D | SA-CNN |
|---------------|---------------------|-----|-----|------|-----|--------|
| Beach         | 87 30 93 93 100 100| 100 | NA  | 97   |
| Elevator      | 87 47 100 97 100 97| 100 | NA  | 100  |
| Forest Fire   | 63 17 67 70 83 93 | 93  | NA  | 100  |
| Fountain      | 43 3 43 57 47 87  | 87  | NA  | 97   |
| Highway       | 47 23 70 93 73 100| 100 | NA  | 97   |
| Lightning Storm| 63 37 77 87 93 97 | 97  | NA  | 97   |
| Ocean         | 97 43 100 100 90 100| 100 | NA  | 100  |
| Railway       | 83 7 80 93 93 100 | 100 | NA  | 100  |
| Rushing River | 77 10 93 87 97 97 | 97  | NA  | 100  |
| Sky-Clouds    | 87 47 83 93 100 97| 97  | NA  | 100  |
| Snowing       | 47 10 87 70 57 97 | 97  | NA  | 98   |
| Street        | 77 17 90 97 97 100| 100 | NA  | 93   |
| Waterfall     | 47 10 63 73 77 83 | NA  | 93  |
| Windmill Farm | 53 17 83 87 93 100| 100 | NA  | 97   |
| Overall       | 68.33 22.86 80.71 | 85.48 | 85.95 | 96.19 | 96.67 | 97.14 |

Table 4: Comparison of classification scores for various methods of Dynamic Scene Recognition on YUPenn dataset.
|          | avalanche | forest fire | landslide | tornado | waves | boiling water | fountain | volcano eruption | whirlpool | chaotic traffic | iceberg collapse | smooth traffic | waterfall |
|----------|------------|-------------|-----------|---------|-------|--------------|----------|-----------------|----------|----------------|-----------------|---------------|----------|
| avalanche | 9 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 | | | | | | | | | | | | |
| forest fire | 0 9 0 0 0 0 0 0 0 1 0 0 0 0 | | | | | | | | | | | | |
| landslide | 1 0 8 0 0 0 0 0 0 0 1 0 0 0 | | | | | | | | | | | | |
| tornado | 0 1 0 9 0 0 0 0 0 0 0 0 0 | | | | | | | | | | | | |
| waves | | | | | | | | | | | | | | |
| boiling water | 0 0 0 0 0 9 0 0 1 0 0 0 | | | | | | | | | | | | |
| fountain | 0 0 0 0 0 0 9 0 0 0 0 0 0 0 | | | | | | | | | | | | |
| volcano eruption | 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 | | | | | | | | | | | | |
| whirlpool | 0 0 0 0 0 0 0 0 0 1 0 0 0 | | | | | | | | | | | | |
| chaotic traffic | 0 0 0 0 0 0 0 1 0 0 0 0 0 0 | | | | | | | | | | | | |
| iceberg collapse | 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 | | | | | | | | | | | | |
| smooth traffic | 0 0 0 0 0 0 0 0 2 0 8 0 | | | | | | | | | | | | |
| waterfall | 0 1 0 0 0 0 0 0 0 0 0 0 0 0 9 | | | | | | | | | | | | |
creases as we take more frames and then saturates roughly after $N > 30$ on both the datasets. From figure 4, we conclude that choosing $N = 30$ is sufficient for getting high accuracy with high confidence. Choosing this optimal value of the number of frames balances out computation time and accuracy.

An interesting thing to note is that, in the case of YUPenn, even for $N = 1$, the accuracy is very high (95%), which is close to the accuracy for larger values of $N$. However, in the case of Maryland, the accuracy improves significantly as $N$ increases. Considering the fact that in Maryland there is lot of camera motion, performing mean aggregation significantly improves the performance as compared to a single frame. But in YUPenn as the camera motion is negligible, even a single frame is very powerful and taking mean results in a slight improvement. This indicates that the simple aggregation scheme robustly handles the effect of camera motion in the videos. We also explored other pooling strategies, such as min and max pooling. While min-pooling in general gave poor accuracy, scores achieved via max-pooling was better. Still, statistical aggregation outperformed these strategies on both the datasets.

We compare our most effective method - statistical aggregation scheme $(\hat{\mu} + \hat{\sigma} + \hat{\gamma} + \hat{\kappa})$ to previous methods for Maryland (Table 3) and YUPenn (Table 4). The proposed approach shows the outstanding performance over current state-of-the-art methods (BoSE & C3D) for the Maryland dataset, with a leap of 13.84%. The classes iceberg collapse and avalanche witness largest improvement over previous best performing techniques. On other classes, the statistical aggregation is either at par or ahead in terms of classification accuracy. The proposed approach also shows marginal improvement in the already saturated dataset YUPenn, as pointed out by recent work [22]. It achieves perfect precision and recall in six out of fourteen classes. Overall, the accuracy of the proposed approach exceeds that of the state of the art methods, BoSE & C3D by 1.04% & 0.47% respectively for YUPenn dataset.
Figure 4: Accuracy vs number of frames using mean aggregation. Here x-axis denotes the number of frames used for aggregation and the y-axis denotes the average accuracy obtained for 18 trials in each case. (a) Maryland (left) and (b) YUPenn (right)

5. Conclusion

As compared to previous spatio-temporal approaches, we focus on capturing temporal variations of very powerful spatial descriptors provided by CNN. This method is computationally efficient than the traditional local feature extraction, encoding and pooling approaches. We observe that CNN spatial descriptors are excellent representatives of spatial information, as demonstrated by the accuracies obtained using only a single frame per video. This is because most of the natural scenes, in spite of the inherent dynamism, are highly correlated with their spatial attributes. However, there is a large uncertainty in the performance, as it is highly dependent on which frame is chosen from the video. We propose methods that utilize multiple frames to improve accuracy as well as reduce this uncertainty to a large extent, as shown in Figure 4.

We evaluate our algorithm on two standard and publicly available datasets (Maryland and YUPenn). Our proposed algorithm shows outstanding performance over the current state-of-the-art methods by 13.84 % for Maryland and 1.05 % for YUPenn datasets. The approach works well even for very challenging Maryland dataset having large camera motion and jitter. High accuracies obtained for the various statistical moments indicate that similar classes have similar temporal statistics and that the spatial features temporally evolve in a similar way. Hence in future, various probabilistic methods can be explored by considering the joint distribution of the 4096 random variables (from the activation features of CNN) and finding out different models for different classes. Expanding the approach to classify datasets with a large number of categories is a challenge which also needs to be investigated.
Acknowledgment

The authors would like to thank SERB-DST for supporting this research through a start-up research grant.

References

References

[1] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, A. Torralba, Sun database: Large-scale scene recognition from abbey to zoo, in: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on, IEEE, 2010, pp. 3485–3492.

[2] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: NIPS, p. 2012.

[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: IEEE CVPR, 2014.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, C3d: Generic features for video analysis, arXiv preprint arXiv:1412.0767.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, arXiv preprint arXiv:1408.5093.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge (2014). arXiv:arXiv:1409.0575.

[7] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: In Workshop on Statistical Learning in Computer Vision, ECCV, 2004, pp. 1–22.

[8] K. Grauman, T. Darrell, The pyramid match kernel: Discriminative classification with sets of image features, in: IEEE ICCV, Vol. 2, 2005, pp. 1458–1465.

[9] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation in: IEEE CVPR, 2010, pp. 3304–3311. URL "http://lear.inrialpes.fr/pubs/2010/JDSP10"

[10] F. Perronnin, C. Dance, Fisher kernels on visual vocabularies for image categorization, in: Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on, 2007, pp. 1–8.
[11] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: IEEE CVPR, 2010, pp. 3360–3367.

[12] C. Schmid, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: IEEE CVPR, 2006.

[13] Y. Gong, L. Wang, R. Guo, S. Lazebnik, Multi-scale orderless pooling of deep convolutional activation features, arXiv preprint arXiv:1403.1840.

[14] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks, CoRR abs/1312.6229.

[15] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 818–833.

[16] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, arXiv preprint arXiv:1310.1531.

[17] A. S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, Cnn features off-the-shelf: an astounding baseline for recognition, arXiv preprint arXiv:1403.6382.

[18] M. Marszalek, I. Laptev, C. Schmid, Actions in context, in: IEEE CVPR, 2009, pp. 2929–2936.

[19] N. Shroff, P. Turaga, R. Chellappa, Moving vistas: Exploiting motion for describing scenes, in: IEEE CVPR, 2010.

[20] A. B. Vasudevan, S. Muralidharan, S. P. Chintapalli, S. Raman, Dynamic scene classification using spatial and temporal cues, in: IEEE ICCVW, 2013, pp. 803–810.

[21] K. G. Derpanis, M. Lecce, K. Daniilidis, R. P. Wildes, Dynamic scene understanding: The role of orientation features in space and time in scene classification, in: IEEE CVPR, 2012, pp. 1306–1313.

[22] C. Feichtenhofer, A. Pinz, R. P. Wildes, Bags of spacetime energies for dynamic scene recognition, in: IEEE, 2014.

[23] E. J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis?, Journal of the ACM (JACM) 58 (3) (2011) 11.

[24] C. Guyon, T. Bouwmans, E.-h. Zahzah, et al., Robust principal component analysis for background subtraction: Systematic evaluation and comparative analysis, Principal Component Analysis, P. Sanguansat, Ed.
[25] Z. Lin, M. Chen, Y. Ma, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, arXiv preprint arXiv:1009.5055.

[26] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 27:1–27:27, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[27] C. Theriault, N. Thome, M. Cord, Dynamic scene classification: Learning motion descriptors with slow features analysis, in: IEEE CVPR, 2013, pp. 2603–2610.

[28] C. Feichtenhofer, A. Pinz, R. P. Wildes, Spacetime forests with complementary features for dynamic scene recognition, BMVC, 2013.

Authors

Aalok Gangopadhyay
Aalok Gangopadhyay is an undergraduate student at Indian Institute of Technology Gandhinagar, majoring in Electrical Engineering. His research interests include Computer Vision and Machine Learning.

Shivam Mani Tripathi
Shivam Mani Tripathi is an undergraduate student at Indian Institute of Technology Gandhinagar, majoring in Electrical Engineering. His research interests span across machine learning and computer vision.

Ishan Jindal
Ishan Jindal is currently a Junior Research Fellow at the Indian Institute of Technology Gandhinagar under Prof. Shanmuganathan Raman. He received his M.Tech. degree from Indian Institute of Technology Roorkee. His research interests are 3D reconstruction, Computer Vision, Image and Video Processing and Machine Learning. He was also awarded DAAD Masters Scholarship for the year 2013-14.
Shanmuganathan Raman
Shanmuganathan Raman is currently an Assistant Professor of Electrical Engineering at the Indian Institute of Technology Gandhinagar. Prior to this position, he was an Assistant Professor at the Indian Institute of Technology Jodhpur. Prior to that he was a Postdoctoral Research Associate at the Indian Institute of Science, Bangalore. He received his B.E. and M. Tech. degrees from PSG College of Technology, Coimbatore and the Indian Institute of Technology Bombay, respectively. He obtained his Ph.D. degree from the Indian Institute of Technology Bombay. His research interests include Computer Vision, Computer Graphics, and Computational Photography. He was awarded Microsoft Research India Ph.D. Fellowship for the year 2007.