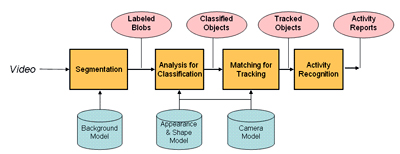
**Deep Learning in GPU**

**Introduction**

Data is nothing but information represented in various forms. It is a set of values of qualitative or quantitative variables which is measured, collected, reported and analyzed. It can be represented in many forms ranging from the traditional row-columns to images and videos. Due to the boom of social media there has been a tremendous increase in the number of images and videos being generated and shared. This rapid increase in multimedia has grabbed the interest of data scientists all over the world in introducing the concept of ‘Semantic Analysis of Multimedia’. The emergence of smart cameras and digital video recorders has encouraged the scope for extraction of meaningful content from videos. This advent was supported with various tools in computer vision, pattern analysis and machine learning. This concept is coined as Video Semantic analysis (VCA) or intelligent video. Among all the above strategies, Machine learning is chosen. Deep Learning is a branch of machine learning and has a wide set of implementation concepts such as Neural Networks. Neural Net henceforth is our chosen construct for VCA. GPU is being used to perform deep learning with videos as the dataset. This survey presents the various techniques and methodologies which are used to derive insights into the video.

**Main Body**

There exists series of processing stages in video analytics. Fundamentally, the changes over the successive frames need to be detected. These changes should be later qualified, correlated and construed over all the frames. The process of detecting changes and extracting relevant changes for further analysis and qualification is called Segmentation. There are two kinds of pixels; Background pixels, the ones which do not change and the Foreground Pixels, the ones which change. Since segmentation subtracts the background pixels to get foreground pixels, it is also called background subtraction. The quantum of change used to identify the foreground pixels is a key factor. Segmentation produces one or more foreground blobs, a blob being a collection of connected pixels. Moving along, classification is done on the blobs. Either a single frame or multiple frames are taken into consideration while classifying. In each frame a particular property or a feature or a key point is extracted which directs us to a certain class. Many such properties are extracted which lets us to decide as to which category it belongs. Sometimes, Recognition entails classification based on the need to extract objects (a collection of features).For the identified features or objects, tracking needs to be done along the frames which distinguish the image and video processing. Finally, Activity recognition is performed that fuses all the above three stages to infer an occurrence of an activity. For instance, a blob corresponding to a person moves towards a blob corresponding to an object and later object blob disappears with the person blob, it should be recognized as the person taking away the object [1].



**Fig 1: VCA pipeline**

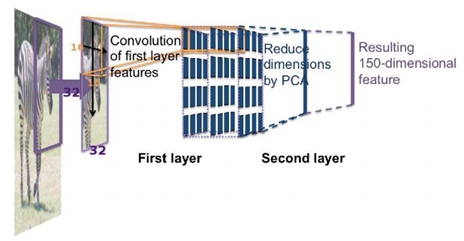
The above pipeline is a generic framework and becomes sophisticated with the desired functionalities. As discussed earlier, there are various strategies that can be deployed in video analysis. One such methodology is Deep Learning. Deep Learning is a branch of machine learning based on a set of algorithms which attempt to model high-level abstractions in data by utilizing the concept of multiple processing layers which have complex structures.

In the line of development of various methodologies for analysis of videos, the first in the line was Support Vector Machines (SVM). SVM is a supervised learning method which creates a hyper plane for classifying the labels. The optimal hyper plane generated as a result of training, categorizes the test sample. Emotion recognition using videos is an application of VCA. In the SVM based training the facial features are extracted initially and a displacement vector is created. Blobs of features together extracted leads to the identification of an object. For instance, key features of eye, ear, nose, mouth, cheeks and forehead is detected as face. The displacement vectors between the initial frame and the peak frame is the main criteria behind emotion recognition. The feature extraction is done using PCA and Eigen features. Michel.et.al used *libsvm* as an underlying SVM classifier. The advantage behind this approach is the support for combination of training set of multiple individuals. However this comes at the cost of complexity in implementation and time inefficiency. The gestures such as head tilting produce inaccuracies [2].

SVM being the traditional approach has other major limitations associated with it as well. It can't handle large data sets, degree of preprocessing is high and the accuracy for a temporal based database is low. This makes us to look for an alternative approach. Hence we bring the light of deep learning into our problem domain. Deep learning particularly CNN has been particularly successful in image based classification. In the case of deep learning the video is preprocessed to generate to frames of the video. These frames are then sent to a Convoluted Neural Network (CNN) for feature extraction. Using the results obtained from the CNN there are 2 ways by which can be performed with respect to video classification. The two options available are: 1) Feature Pooling- Independently process each frame by using a CNN and the result of which is max-pooled. There are many feature pooling methods such as conv pooling, late pooling, slow pooling, local pooling and time-domain convolution. 2) RNN with LSTM- RNN is a Neural Network which is convoluted and contains the additional feature that it maintains previous state of the iteration. Using the saved previous state we can find relations between the current and previous state. The drawback of this approach is the emergence of exploding and vanishing gradients. The exploding gradient is resolved by using the fixing the upper bound/limit. The vanishing gradient is resolved by using the concept of LSTM (Long-Short Term Memory). The LSTM contains 3 gates input, output and forget gate which are used to control the write, read and delete operations to the memory cell respectively. The advantage is using a LSTM is that the memory can be preserved over a long period of time. The above proposed approach of Deep Learning is more accurate for video data when compared to the traditional approach of a svm classifier. The degree or level of preprocessing is comparatively less in the deep learning approach compared to other approaches [14].

There are the concepts of physics which are borrowed to VCA as well. One among them is the concept of temporal coherence. Classical learning and transduction can handle only labeled data; thereby the unlabeled data is completely obliterated. There is pivotal information contained in unlabelled data as well which can be exploited by temporal coherence. Mobahi.et.al in his work uses two successive frames which are likely to contain similar contents and represent same concept classes with regard to small transformations like rotation, deformation or translation. This similarity is Temporal Coherence. A deep Convolution Neural Network that performs chain of filters and resolution reduction steps is chosen for object recognition. This is advantageous over Transductive Support Vector Machines (TSVMs) as there is no assumption that unlabelled videos should belong to the training set and does not rely on low density assumption as that in graph based approaches. COIL100 was the dataset used for this implementation and versatility of the model was enhanced by using similar images as that of COIL100. An accuracy of 71.49% was obtained for labeled examples of COIL100. An improvement of 7% and 8% on Animal dataset and COIL100 like dataset was observed compared to plain CNN without using unlabelled video. This suggests that improvements can be achieved when unlabeled data is from the same dataset as the labeled data and the use of unlabeled data is favorable even when the sets are from different datasets [3].

Another work which is performed on unsupervised learning is using the concept of Slow Feature Analysis. Zou.et.al deals with tracking and feature detection in videos that simulate fixations and smooth pursuit in human vision. This feature is spatial but not spatial temporal and is suited for extracting features from still images. The datasets used in this approaches are COIL100, CALTECH101, STL-10 and PubFig. The system uses two techniques: Fixed location video sequences and video sequences obtained with tracking.

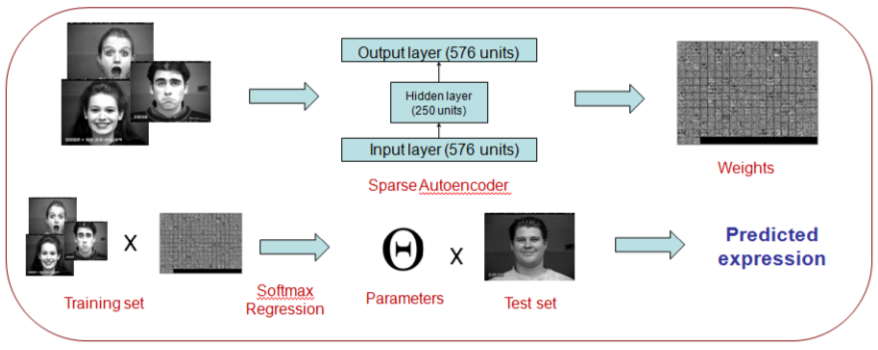


**Fig 2: Architecture used to learn invariance from videos**

Emphasizing on the fundamental concept, Slow Feature Analysis (SFA) is used for mapping data into quadratic equation and performing eigenvector decomposition, which optimizes temporal slowness. For learning invariant features from video, we use unsupervised learning algorithm using temporal slowness principle. Invariance visualization method has 2 layer pooling wherein image of size 16\*16 is used in one layer and image of 32\*32 is used in the other. The sequences in a video can be tracked by using Normalized cross correlation (NCC). In self-taught learning framework, system primarily learns set of features from unlabeled videos using simulated fixations, and then applies the features learnt to classify tasks. This improves the accuracy by 4% to 5% across all the four image recognition datasets [11].

Considering the field of sports, there is wide scope of semantic extraction. Recognizing the various tricks and actions performed is a pivotal task. Zhang.et.al proposes few approaches for analysis of event detection and semantics extraction, video analysis for event structure modeling and event moment detection, text or video alignment for event boundary detection in the video. An external resource is used i.e., web-casting texts, to extract semantics and detect events from basketball games. There are three major improvements: a) Improvement in the game time recognition algorithm. b) Need for an unsupervised clustering method to automatically detect event from web-casting text instead of predefined keywords. c) Usage of statistical approach instead of finite state machine to detect event boundary in the video. The framework contains 4 major parts: a) web-casting text analysis: description of the game progress as in exciting or important events that happened in that game using certain predefined key-words. b) Broadcast video analysis: visualize the video stream to detect event motion carried out using time-tagged events. c) text/video alignment : the mapping of time-tagged text event to game video stream using Hidden Markov Model (HMM) to model the temporal event structure and detect the event boundaries. The shot containing the detected event is used as the reference shot to obtain a search range by a rule for event boundary detection. d) Semantic annotation and indexing for personalized retrieval [13].

Borrowing the concepts of deep networks, a slightly modified network was created. It was named as Deep Belief network. Deep Belief Networks (DBNs) and sparse autoencoder was used to recognize facial expressions from videos, as DBN involves human perception such as touch, vision, audio etc, while sparse autoencoder will remove the necessity to perform feature selection separately by activating different neurons for each expression. Rao.et.al uses the AU-Coded Cohn Kanade Facial expression dataset. The best set of weights was picked for autoencoder and was trained using only the feature activations as input to softmax layer, an accuracy of 22% was observed. The accuracy was greatly improved with the use of logistic regression to individual expressions [6]. The entire methodology is delineated in Fig 3.

**Fig 3: Methodology incorporated for feature extraction and classification**

When we consider the application of Deep learning to the current growing society, the first thing which pops in our mind is the idea of Smart city. The ongoing activities of a city are captured through sensors which consists of data such as Transportation, Video surveillance etc. Deep Learning is apt for processing and analyzing such large-scale data and Wang.et.al provides the following information. Deep Learning architectures have different variants such as Convolutional Neural Networks (CNN), Deep Belief Networks (DBN), Deep Boltzmann Machines (DBN) etc. Also the use of Graphics Processor Unit (GPU), an electronic circuit designed for accelerating algorithms is contributing to the success of Deep Learning. It is also observed that GPU speedup on learning DBNs achieves upto 70 times against dual-core CPU implementation. The tasks performed aid in building a smart city to its utmost sophistication [8].

Narrowing down the societal impacts of VCA, we come across pedestrian detection. Video based pedestrian detection is one of the trending subjects in computer vision due its direct application in surveillance and automotive safety applications. Taking a deviation from deep learning implementations, we can throw light on the other methodologies followed. Cho.et.al tries to develop a real-time pedestrian detection system for both manually-driven and autonomous vehicles. Caltech pedestrian dataset is used for this work. The detection system algorithm is implemented in C to optimize several subsystems and improve overall speed. The geometric constraint analysis based on known camera calibration is used to enhance efficient search and improve detection accuracy and quantitative analysis of the system with public real world datasets is carried out. This system gives a detection rate of 61% with 1 false positive per image (FPPI). The detection accuracy can be further increased by using partial occlusion handling algorithms which are in the process of development. One clear aspect what can be noticed is that there is no deep learning concepts used in this work. This shows that for Video analysis, Computer vision supports other approaches as well but has its disadvantages entailed. Since Deep learning is growing rapidly, it would be an effective tool in video analysis [10].

Getting back to the sea of Deep learning, we have visualized that CNN has proved its remark in analyzing the videos. Yet another sophisticated neural network called Recurrent Neural Network (RNN) also creates a benchmark in VCA. RNN helps to create a dynamic temporal behavior where in which units are connected in the form of directed acyclic graph. The added advantage of RNN is that it can also perform generation apart from prediction. A simple RNN model having a compelling limitation that strictly integrate state information over time is known as “vanishing gradient” effect generally defined as the ability to back propagate an error signal through a long-range temporal interval which becomes increasingly impossible in practice, that are generally applicable to visual time-series modeling. Donahue.et.al deals with 3 main tasks: Activity recognition, Image description and video description. UCF-101 dataset is used for all the 3 tasks.

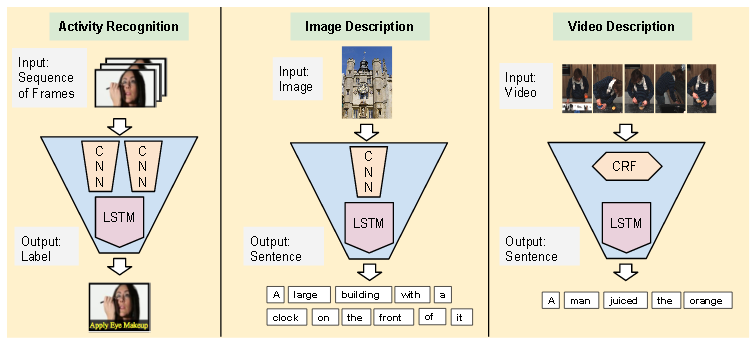
VISUAL FEATURES

SEQUENCE LEARNING

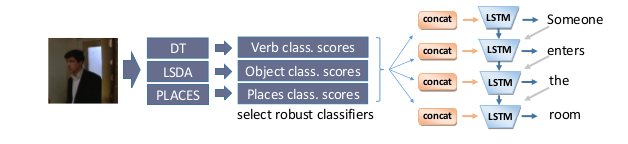
PREDICTIONS

VISUAL INPUT

**Fig 4: Proposed Long-term Recurrent Convolutional Networks (LRCNs)**

**Fig 5: Task specific instantiations of LRCN model**

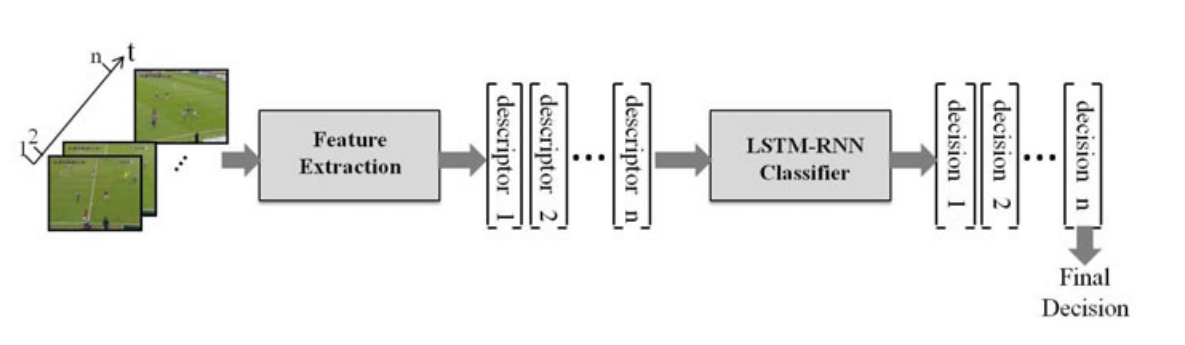
Activity description: Here, N individual frames are given as input into N CNNs which are connected to a single-layer LSTM with 256 hidden units. This gives the description of the activity carried out by the sequence of frames. This gives an accuracy of about 60.2% for hybrid model and 57.4% for Caffe reference models. There are two tasks performed mainly Image Description and Video Description. The Image description task is a static process that requires a single convolution network, since the input consists of single frame. This process utilizes VANILLA RNN and training datasets, a combination of Flickr30k and COCO2014 to evaluate the sentence generation, and they also use BLEU metric and Amazon Mechanical Tuskers to evaluate generated sentences. The objective of video description is to generate variable length stream of words that describe the video. The approaches carried out in this work for video description are: a) LSTM encoder and decoder with CRF max. b) LSTM decoder with decoder CRF max. c) LSTM decoder with CRF prob [9].

Moving along with the other datasets, we have the movie description database. Two large scale movie description datasets namely MPII Movie Description (MPII-MD) and Montreal Video Annotation Dataset are associated with textual descriptions and studying of how to generate movie description for visually disabled people. Rohrbach.et.al contributes a) an approach to build visual classifiers to distinguish verbs, objects, and places extracted from weak sentence annotations. b) based on the visual classifiers the evaluation of different design choices to train a Long Short term Memory (LSTM) for generating descriptions, which outperforms previous works on the MPII-MD dataset. c) A detailed analysis of prior work and current approach to understand the challenges of the movie description task. The overview of the approach is shown in Fig 6.

**Fig 6: Overview of the approach**

LSTM as described earlier is a type of RNN based architecture which overcomes the problem of vanishing and exploding gradients. It has various units such as input, output and forget gates which decide the overall cycle of data through the LSTM structure. The approach followed is similar to the one proposed by Hendricks et al., which uses LSTM implementation based on Caffe. The movie description is carried out in a two step approach where the first step performs visual recognition with the use of classifiers trained according to the labels, and the second step is the textual description generation using LSTM network and classifier scores. This approach identifies activities, objects and places better than related work such as SMT and S2VT. This method also performs better on longer sentences and on sentences that are difficult to retrieve [7].

Initially the semantic event extraction was performed on the basketball videos, but here an LSTM and RNN coupled action classification is performed on the soccer videos (MICC-Soccer-Actions-4 database). The major challenge in information systems is automatic video indexing. The work by Baccouche.et.al basically deals with the action classification like Shot-on-goal, Placed-kick, Throw-in and Goal-kick experimented on the MICC-Soccer-Actions-4 database. A set of features like visual content by means of Bag of Words (BoW) approach and the dominant motion by a key point based approach are described. The approach is shown in the Fig 7.



**Fig 7: The Proposed Classification Scheme**

The soccer video is divided into sequences and each sequence is represented by a descriptor corresponding to a set of features. The successful classification depends upon on these features. A Recurrent Neural Network (RNN) along with Long Short-Term Memory (LSTM) neurons is trained to classify each action type based on the temporal evolution of the descriptors. BoW approach for visual content representation: The image is represented by means of histogram of visual words to a set of local features extracted from the image. Based on the associated words/descriptors, it is compared with codebook containing the values that encode frequency of words associated with that sequence. A SIFT-Based Approach for Dominant Motion Estimation: The assumption made here is that the movement of camera is affine and estimate the affine transformation between image Lt at time t and image Lt+1 at time t+1 to detect dominant motion. When the above 2 mentioned steps are done, then next step is to classify the action in the given video using LSTM-RNN. Here, one hidden layer in LSTM is used and the input layer has variable size depending on which features are fed. In the output layer, softmax function is used so that the outputs are between 0 and 1. The combination of the two features gives a classification rate of 92% [12].

**Conclusion**

The above presented works clearly depicts the work performed in getting insights into the videos. May it be detection, recognition or prediction of classes, deep learning has proved its efficiencies with its wide variety of constructs. Taking into account of all the above approaches and their feasibilities CNN based approach is the one which grabs our interest. The MMI dataset [4] and the MIT Affectiva dataset [5] are the popular datasets used for emotion detection. These datasets are preprocessed and effectively engrossed to detect the emotions and the expressiveness respectively. Emotion Detection aids in Behavioral Analytics where the current research is going on. Traditional text mining and sentimental analysis algorithms are used to determine the positivity of an individual, where as a video based emotion detection is a much more sophisticated and better approach in achieving the goal. This gives further scope for various insights to be drawn and has various applications across all the fields.

**References:**

1. <http://www.eetimes.com/document.asp?doc_id=1273834> : EE-Times Webpage regarding video analytics.
2. Philipp Michel and Rana el Kaliouby, “Real Time Facial Expression Recognition in Video using Support Vector Machines.”, Proceedings of the 5th International Conference on Multimodal Interfaces, Vancouver, Canada, 2003.
3. Hossein Mobahi, Ronan Collobert and Jason Weston, “Deep Learning from Temporal Coherence in Video”, Proceedings of the 26th Annual International Conference on Machine Learning, ICML ’09. ACM, 2009.
4. Maja Pantic, Michel Valstar, Ron Rademaker and Ludo Maat, “Web-based Database for Facial Expression Analysis”, Proceedings of the IEEE International Conference of Multimedia and Expo, 2005.
5. McDuff, Daniel Jonathan; el Kaliouby, Rana; Senechal, Thibaud; Amr, May; Cohn, Jeffrey F.; Picard, Rosalind W. “Affectiva-MIT Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected In-the-Wild.” Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2013.
6. Adithya Rao and Narendran Thiagarajan, “Recognizing facial expressions from videos using Deep Belief Networks”, Stanford CS 229 Machine Learning Final Projects, Technical Report, 2010.
7. Anna Rohrbach, Marcus Rohrbach and Bernt Schiele, “The Long-Short Story of Movie Description”, Proceedings of the German Conference on Pattern Recognition, 2015.
8. Li Wang and Gang Wang, “Deep Learning Algorithms with applications to Video Analytics for a Smart City: A Survey”.
9. Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko and Trevor Darrell, “Long-term Recurrent Convolutional Networks for Visual Recognition and Description”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
10. Hyunggi Cho, Paul E. Rybski, Aharon Bar-Hillel and Wende Zhang, “Real-time Pedestrian Detection with Deformable Part Models”, Proceedings of the IEEE Intelligent Vehicles Symposium, 2012.
11. Will Y. Zou, Shenghuo Zhu, Andrew Y. Ng and Kai Yu, “Deep Learning of Invariant Features via Simulated Fixations in Video”, Proceedings of the Advances Neural Information Processing System Conference, Lake Tahoe, 2012.
12. Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia and Atilla Baskurt, “Action Classification in Soccer Videos with Long Short-Term Memory Recurrent Neural Networks”, Proceedings of the International Conference of Artificial Neural Networks (ICANN), 2010.
13. Yifan Zhang, Changsheng Xu, Yong Rui, Jinqiao Wang and Hanqing Lu, “Semantic Event Extraction From Basketball Games Using Multi-Modal Analysis”, Proceedings of the IEEE International Conference of Multimedia and Expo, 2007.
14. Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga and George Toderici, “Beyond Short Snippets: Deep Networks for Video Classification”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.