

# ACTION RECOGNITION IN HAZE

P. Prathyush  
CSE  
IIIT Sri City  
prathyush.p16@iiits.in

K. Sai Suhas Tanmay  
ECE  
IIIT Sri City  
saisuhashtanmay.k16@iiits.in

Tanneru Sri Girinadh  
ECE  
IIIT Sri City  
srigirinadh.t16@iiits.in

B.S.N.V. Chaitanya  
ECE  
IIIT Sri City  
viswachaitanya.b16@iiits.in

Anjali Poornima K  
ECE  
IIIT Sri City  
anjaliipoornima.k16@iiits.in

**Abstract**— Action recognition in video sequences is a challenging problem of computer vision due to the similarity of visual contents, changes in the viewpoint for the same actions, camera motion with action performer, scale and pose of an actor, and different illumination conditions. Also there is no designated action recognition model for hazy videos. This paper proposes a novel unified and unique model for action recognition in haze built with Convolutional Neural Network(CNN) and deep bidirectional LSTM (DB-LSTM) network. First, every frame of the hazy video is feeded into the AOD-Net(All-in-One Dehazing Network). Next, deep features are extracted from every sampled dehazed frame by using VGG-16, which helps reduce the redundancy and complexity. Later, the sequential and temporal information among frame features is learnt using DB-LSTM network, where multiple layers are stacked together in both the forward and backward passes of DB-LSTM to increase its depth. The proposed unified method is capable of learning long term sequences and can process lengthy videos(even hazy videos) in real time by analyzing features for a certain time interval. Experimental results on both synthesized and natural video datasets show decent results on par with other state of the art methods in action recognition using the proposed method on the benchmark data set UCF-101.

**Keywords**—CNN, Bidirectional LSTM, Haze, Deep Learning

## I. INTRODUCTION

Outdoor photography often suffer from bad weather conditions, observed objects lose visibility and contrast due to the presence of atmospheric haze, fog, and smoke. Haze and fog dramatically degrades the visibility of outdoor images, where contrasts are reduced and surface colors become faint. Moreover, a hazy video will put the effectiveness of many subsequent high-level computer vision tasks in jeopardy, such as object detection and action recognition. There is a series of image degradation in the video acquired in haze and other weather. Virtually all computer vision tasks or computational photography algorithms assume that the input images are taken in clear weather. Unfortunately, this is not always true in many situations, therefore dehazing is highly desired. For these applications, removing haze for the input videos will be a useful pre-processing. However, removal of haze is a challenging and complex problem as the haze is dependent on the depth information which is not available due to visual degradation. In the context of videos, an action is represented using a sequence of frames, which humans can

easily understand by analyzing contents of multiple frames in sequence.

## II. RELATED WORKS

Various methods have been proposed which use additional information other than depth for dehazing the image. Methods based on Depth [5, 11] require the depth information from the user inputs or known 3D models. Recently, single image haze removal[2, 16] also has made significant and decent progress. A stronger prior or assumption will make these methods successful. In DCP[], a simple but effective image prior - dark channel prior is proposed to remove haze from a single input image.. It is based on a key observation - most local patches in haze-free outdoor images contain some pixels which have very low intensities in at least one color channel. A high quality depth map can also be obtained as a by-product of haze removal. However, this approach cannot well handle heavy haze images and fails in the cases where the assumption breaks.

All the work carried in the image dehazing focused on the classical atmospheric scattering model:

$$I(x) = J(x)t(x) + A(1 - t(x)) \quad (1)$$

where  $I(x)$  is observed hazy image,  $J(x)$  is the scene radiance (“clean image”) to be recovered.  $A$  denotes the global atmospheric light, and  $t(x)$  is the transmission matrix.

$$t(x) = e^{-\beta d(x)} \quad (2)$$

$$J(x) = \frac{1}{t(x)}I(x) - A\frac{1}{t(x)} + A. \quad (3)$$

Later many CNN based methods (Cai et al. 2016; Ren et al. 2016) employ CNN as a tool to regress  $t(x)$  from  $I(x)$ . With  $A$  estimated using some other empirical methods, they are then able to estimate  $J(x)$  by (3).

. Lately AOD-NET(All-in-One Dehazing Network) has a complete end-to-end CNN dehazing model based on re-formulating (1), which directly generates  $J(x)$  from  $I(x)$  without any other intermediate step:

$$J(x) = K(x)I(x) - K(x)$$

$$K(x) = \frac{\frac{1}{t(x)}(I(x) - A) + A}{I(x) - 1}. \quad (4)$$

the AOD-Net architecture is composed of two modules: a K-estimation module consisting of five convolutional layers to estimate  $K(x)$  from  $I(x)$ , followed by a clean image generation module to estimate  $J(x)$  from both  $K(x)$  and  $I(x)$  via (4). All those above mentioned methods are designed for single-image dehazing, without taking into account the temporal dynamics in video. When it comes to video dehazing, a majority of existing approaches count on post processing to correct temporal inconsistencies, after applying single image dehazing algorithms frame-wise. Action recognition using deep networks is developed through 3D convolutional kernels which are applied on video frames in a time axis to capture both temporal information and spatial information. Their approach can capture motion and optical flow information because frames are connected by fully connected layers at the end. A multi-resolution CNN framework for connectivity of features in time domain is proposed by [21] to capture local spatio-temporal information. This method is experimentally evaluated on a new ‘‘YouTube 1 million videos dataset’’ of 487 classes. Their recognition rate on UCF101 is 63.3%, which is still too low for such important task of action recognition. Recently, for wide range of tasks like style transfer(Chen et al. 2017), super-resolution(SR)(Kappeler et al. 2016), deblurring(Su et al. 2016) and classification(Karpathy et al. 2014; Shen et al. 2016), there is a growing interest in modeling video using CNNs. Also attempts are made by (Karpathy et al. 2014; Shen et al. 2016), both try different connectivity options for video classification. (Liu et al. 2017) proposes a more flexible formulation by placing a spatial alignment network between frames. A CNN trained end-to-end model is given by (Su et al. 2016) to learn accumulating information across frames for video deblurring.

### III. METHODOLOGY

#### A. AOD-NET:

We have used AOD-NET to dehaze the video. AOD-NET consists of a K estimation module which has 5 convolution layers to estimate  $K(x)$  and many element wise multiplication layers and addition layers to recover the clean image. The K-estimation module consists of five convolution layers, which has ‘‘concat1’’ layer which concatenates features from the layers ‘‘conv1’’ and ‘‘conv2’’. Similarly, ‘‘concat2’’ concatenates those from ‘‘conv2’’ and ‘‘conv3’’; ‘‘concat3’’ concatenates those from ‘‘conv1’’, ‘‘conv2’’, ‘‘conv3’’, and ‘‘conv4’’. The need for using K-estimate module is for complete end-to-end modeling for restoring clean image. One of the important reasons to use AOD-NET is because it can be seamlessly embedded with other deep models, to constitute one pipeline that performs high-level tasks on hazy images, with an implicit dehazing process. Fig 1 shows the different types of convolution layers present in AOD-NET. To this model we give 5 sampled frames of video as input and these frames are then dehazed by the network.

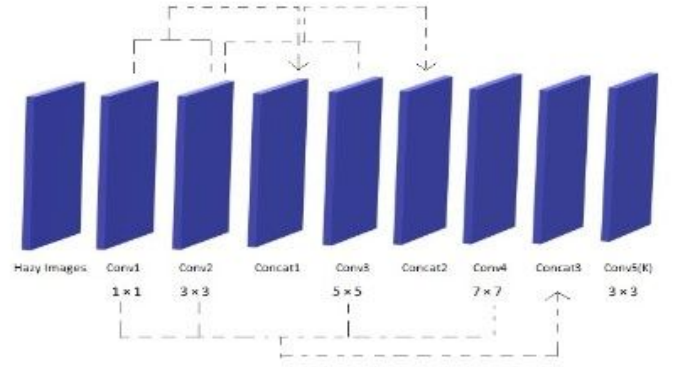


Fig1. AOD-NET Architecture.

#### B. VGG-16:

In traditional architectures like VGG, each successive layer detects features at some more abstractly semantic level than the layer below. VGG-16 is better than any other neural network for feature extraction because the kernel size is less and many convolutional layers are used which gives better results compared to networks with large kernel size and less layers. These dehazed frames are then fed into the VGG-16 network for feature extraction.

#### C. Bidirectional LSTM

Recurrent Neural Networks analyze hidden sequential patterns in both temporal sequential and spatial sequential data. The disadvantage with RNN is that as the time steps increase, it fails to derive context from time steps which are much far behind. Therefore, RNN is able to remember only short-term memory sequences. To solve this problem, Long Short-Term Memory(LSTM) networks are used. LSTM networks are capable of learning long term dependencies. They consist of various gates such as input, output, and forget gates which control the long term sequence patterns. Bidirectional LSTM have two RNNs stacked on top of one another out of which one RNN goes in the forward direction and another one goes in the backward direction. The combined output is then computed based on the hidden state of both RNNs. In our proposed model, we use multiple LSTM layers, so our model has two LSTM layers for both forward and backward passes. After features are extracted, the feature vectors are feeded to the bidirectional LSTM network which then outputs the temporal and spatial interpretations. These are then sent as inputs to a softmax layer. The softmax layer outputs the probabilities for each class and the class with the highest probability is the predicted action in the video. For training, we first take each frame of the video and extract the features from it. Similarly features are extracted from every frame in the video and the whole stack of features are saved into a npy file. Hence in the end we have all the features of the videos in npy files.

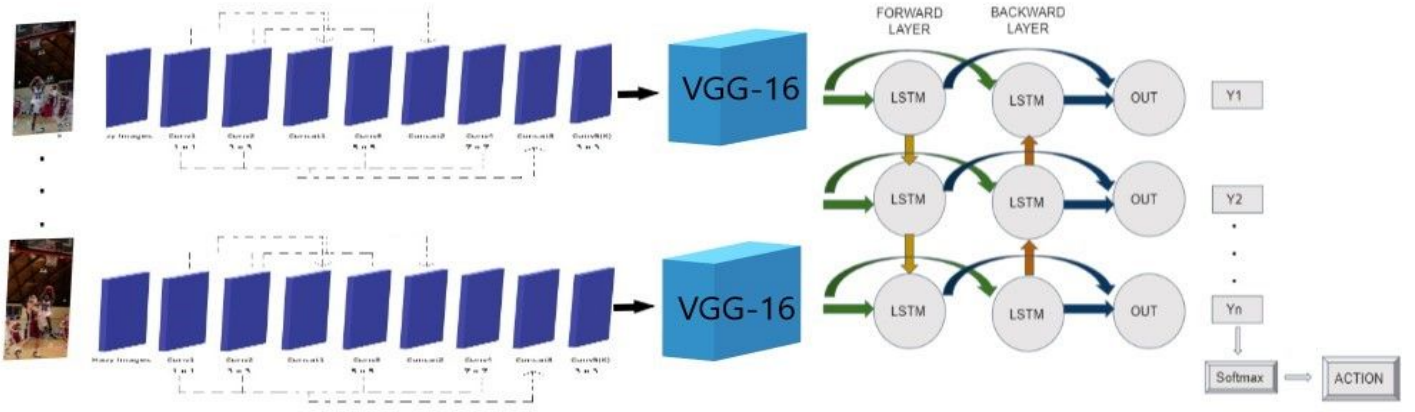


Fig 2. Unified Model Architecture

#### D. Dataset

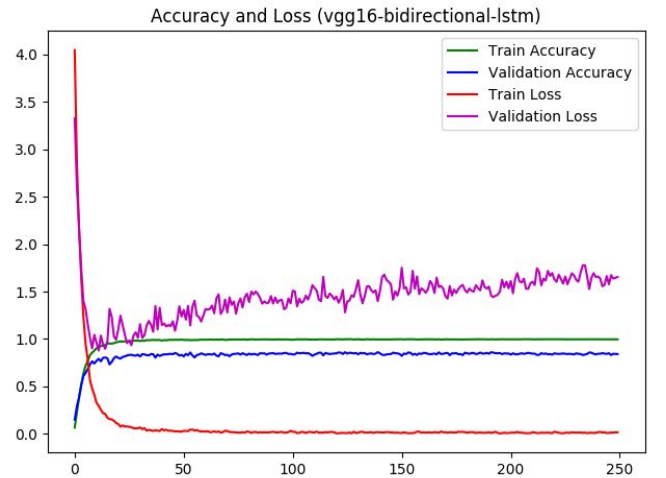
The UCF101 dataset is one of the largest dataset of human actions. It consists of 101 action classes, over 13k clips and 27 hours of video data. The database consists of realistic user uploaded videos containing camera motion and cluttered background. Additionally, it contains baseline action recognition results on this new dataset using standard bag of words approach with overall performance of 44.5%. To the best of our knowledge, UCF101 is one of the most challenging dataset of actions due to its large number of classes, large number of clips and also unconstrained nature of such clips. Due to the lack of hazy video dataset available, we create our own synthesized hazy video dataset by (1), using the ground-truth images with depth meta-data from the UCF101 dataset. For every second frame in the clear video, we calculate the depth map by taking it as a stereo-image based on the assumption that there won't be any significant difference between two continuous frames as we are running them on code snippets. This is anyway done only during the training phase and not during prediction phase and hence there is no problem with the above assumption. Now that we have calculated the depth maps, the transmission map can be obtained by using (2). Next,  $I(x)$  is calculated by using (1). The  $I(x)$  calculated is our hazy image. The above process is repeated for each frame of the clear video and finally, a hazy video is obtained. Therefore we finally generate a synthetic hazy video dataset consisting of 98 videos each for the 101 classes in the UCF101 dataset.

#### IV. EXPERIMENTAL EVALUATION AND RESULTS

The proposed model is tested on both the UCF101 natural dataset and the hazy synthetic UCF101 dataset. Table I. shows the predictions of the proposed model for action recognition for sample clips on both clear and the synthetic hazy datasets of UCF101. In Table I, row 2 contains images correctly classified from the clear UCF101 dataset. Row 3 contains images wrongly classified from the clear UCF101 dataset. Row 4 contains images correctly classified from the hazy synthetic UCF101 dataset. Row 5 contains images wrongly

classified from the hazy synthetic UCF101 dataset. These incorrect predictions are due to the similarity of visual content, motion of camera, and changes in parts of an actor body in both action categories. We have jumped 5 frames in overall experiments because of its optimal results in complexity and accuracy. The proposed method is evaluated on Tesla K80 GPU for feature extraction, training, and testing. The system takes approximately 0.24 sec for feature extraction per frame. Feeding the extracted features to DB-LSTM for classification takes 0.70 sec for 30 frames per second video clip. Overall, the proposed method takes approximately 1.30 seconds for processing of a 1-second video clip. That is our method can process about 30 frames per second, making it a suitable candidate for action recognition in real-time video processing applications on par with other state of the art models.

For training, the proposed method is trained on 85% videos of both datasets, clear and hazy synthetic UCF101 datasets.. 250 epochs were done on the hazy UCF101 dataset.



#### V. Conclusion and Future Work

In the paper, we proposed an action recognition in haze framework which first dehazes the sampled frames, then learns the features and is then fed into the DB-LSTM for classification of action. The proposed model is the first of its kind with unified end-to-end modeling.



Sampled frames are led

After CNN features are extracted from the dehazed video frames, they are fed into DB-LSTM, where two layers are stacked on both forward and backward pass of the LSTM. This helped in recognizing complex frame to frame hidden sequential patterns in the features.

The experimental results indicate that the recognition score of the proposed method gives exceptional results on UCF-101 datasets.

These characteristics make our proposed method more suitable for processing of visual data and can be an integral component of smart systems. The proposed method extracts features from the whole frame of the video. In future, we aim to analyze only the salient regions of the frames for action recognition. Furthermore, we have intention to extend this work for activity recognition in videos [42]–[44]. Finally, the proposed method can be combined with people counting

Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.





- Identify applicable funding agency here. If none, delete this text box.
- Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (bullet list)

A. Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate

TABLE I. PREDICTIONS OF THE PROPOSED MODEL FOR ACTION RECOGNITION FOR SAMPLE CLIPS.

Frames(Ground Truth - Prediction)	Case
	Clear images with correct prediction
	Clear images with wrong prediction
	Hazy images with correct prediction
	Hazy images with wrong prediction

equations with commas or periods when they are part of a sentence, as in:

$ab \bullet \blacksquare$  . !

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## B. Some Common Mistakes

- The word “data” is plural, not singular.
- The subscript for the permeability of vacuum  $\mu_0$ , and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
- Do not use the word “essentially” to mean “approximately” or “effectively”.
- In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
- Do not confuse “imply” and “infer”.
- The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the “et” in the Latin abbreviation “et al.”.
- The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

## V. USING THE TEMPLATE

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

### A. Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future

citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

1) *For papers with more than six authors:* Add author names horizontally, moving to a third row if needed for more than 8 authors.

2) *For papers with less than six authors:* To change the default, adjust the template as follows.

a) *Selection:* Highlight all author and affiliation li

b) nes.

c) *Change number of columns*: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

d) *Deletion*: Delete the author and affiliation lines for the extra authors.

### B. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named "Heading 1", "Heading 2", "Heading 3", and "Heading 4" are prescribed.

### C. Figures and Tables

a) *Positioning Figures and Tables*: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I. TABLE TYPE STYLES

Table Head	Table Column Head		
	Table column subhead	Subhead	Subhead
copy	More table copy <sup>a</sup>		

<sup>a</sup> Sample of a Table footnote. (Table footnote)

Fig. 1. Example of a figure caption. (figure caption)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and

units. For example, write "Temperature (K)", not "Temperature/K".

### ACKNOWLEDGMENT (Heading 5)

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

### REFERENCES

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

- [1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)
- [2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
- [4] K. Elissa, "Title of paper if known," unpublished.
- [5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
- [6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetism Japan, p. 301, 1982].
- [7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord "Format" pull-down menu, select Text Box > Colors and Lines to choose No Fill