BUBBLENET:

A disperse recurrent structure to recognize activities

Igor L. O. Bastos
igorlobastos@dcc.ufmg.br

Victor H. C. Melo
victorhcmelo@dcc.ufmg.br

William R. Schwartz
william@dcc.ufmg.br

Smart Sense Laboratory
Federal University of Minas Gerais - UFMG

DCC
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

ACTIVITY RECOGNITION

Draw Sword: 0.62
Fencing: 0.38

Riding Horse: 0.83
Riding Bike: 0.17

- One of the **most studied topics** in computer vision [1], presenting several **challenges** and **high applicability** for several tasks

- **Discriminative information** denoted by changes in **appearance** and **motion** of elements over time [2, 3]

[1] T. Xu and E. Wong, "Learning temporal structures for human activity recognition," in IEEE BMVC, 2017.
[2] J. Chaquet, E. Carmona, and F. Antonio, "A survey of video datasets for human action and activity recognition," Computer Vision and Image Understanding, 2013.
[3] I. Bastos, L. Soares, andW. Schwartz, "Pyramidal zernike over time: A spatiotemporal feature descriptor based on zernike moments," in CIARP, 2017.

ACTIVITY RECOGNITION

Basal Segments:

Frontal Hit

Parry Sword

Lack of global order:

Frontal Hit + Parry Sword = Fencing

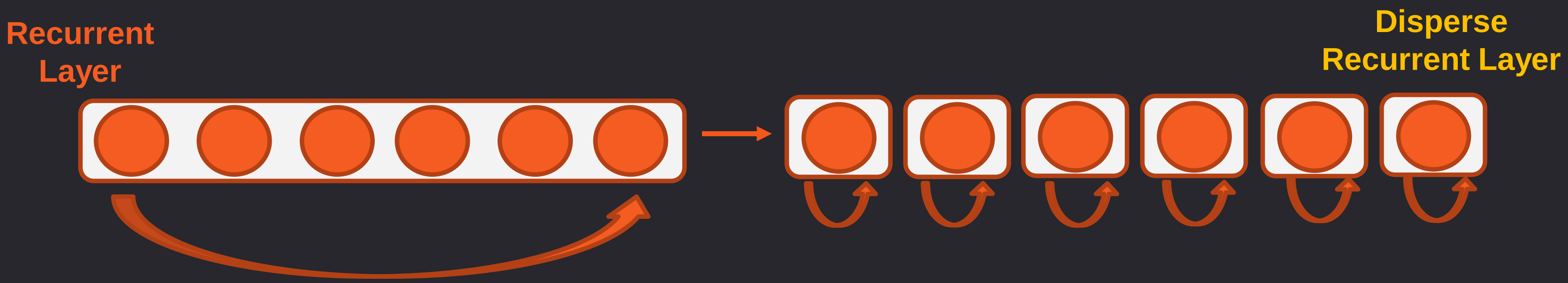
Parry Sword + Frontal Hit = Fencing

- Complex activities can be decomposed into **correlated primitive segments**, being considered their **fundamental parts**, such as **local motions** and **sub-actions** [2, 4, 5]

- Despite the existence of local arrangement, activities **do not present a global temporal structure** of their segments

[4] H. Wang, A. Klser, C. Schmid, and C. Liu, “Dense trajectories and motion boundary descriptors for action recognition,” International Journal of Computer Vision, vol. 103, 05 2013.
[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in IEEE ICCV, 2015.

BUBBLENET



- Literature approaches **do not invest** on the modeling **of correlated fundamental segments** and their **disposition** along with activity videos [5, 6, 7]

- BubbleNET tackles these gaps, through a **recurrent layer** dispersed into **independent modules**, named **bubbles**

- BubbleNET is designed to gather both **elementary** and **spread patterns** existent on the **overall composition** of an action

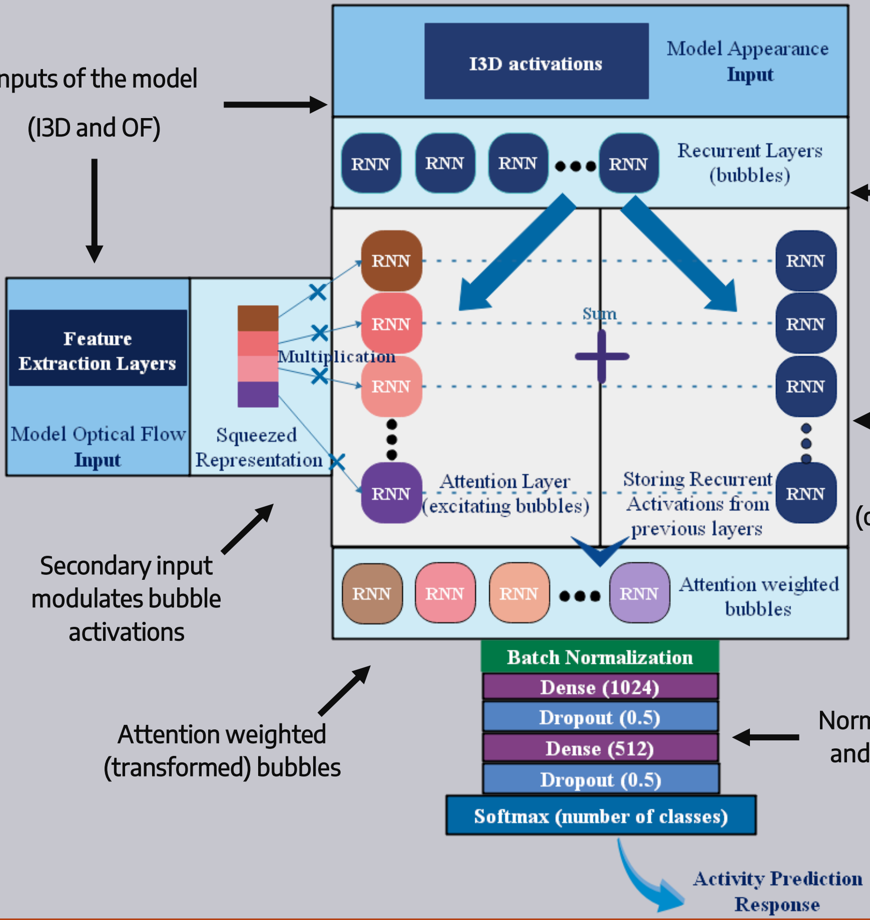
[6] J. Donahue, L. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 677691, 2017.
[7] A. Murad and J. Pyun, "Deep recurrent neural networks for human activity recognition," Sensors, vol. 17, 2017.

BubbleNET Architecture

Original and modulated information are added to produce weighted attention bubbles

Inputs of the model

(I3D and OF)



Dispersed

Recurrent Layer

Bubbles

duplicated

and added

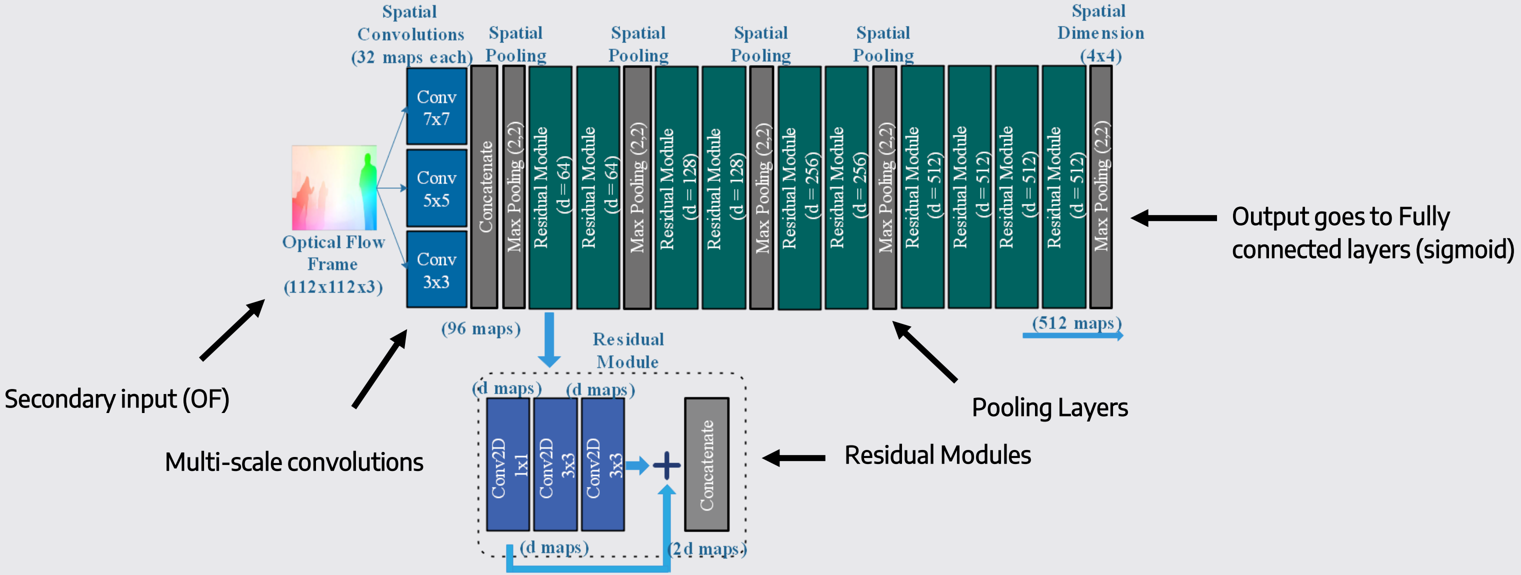
(original and modulated)

Secondary input modulates bubble activations

Attention weighted (transformed) bubbles

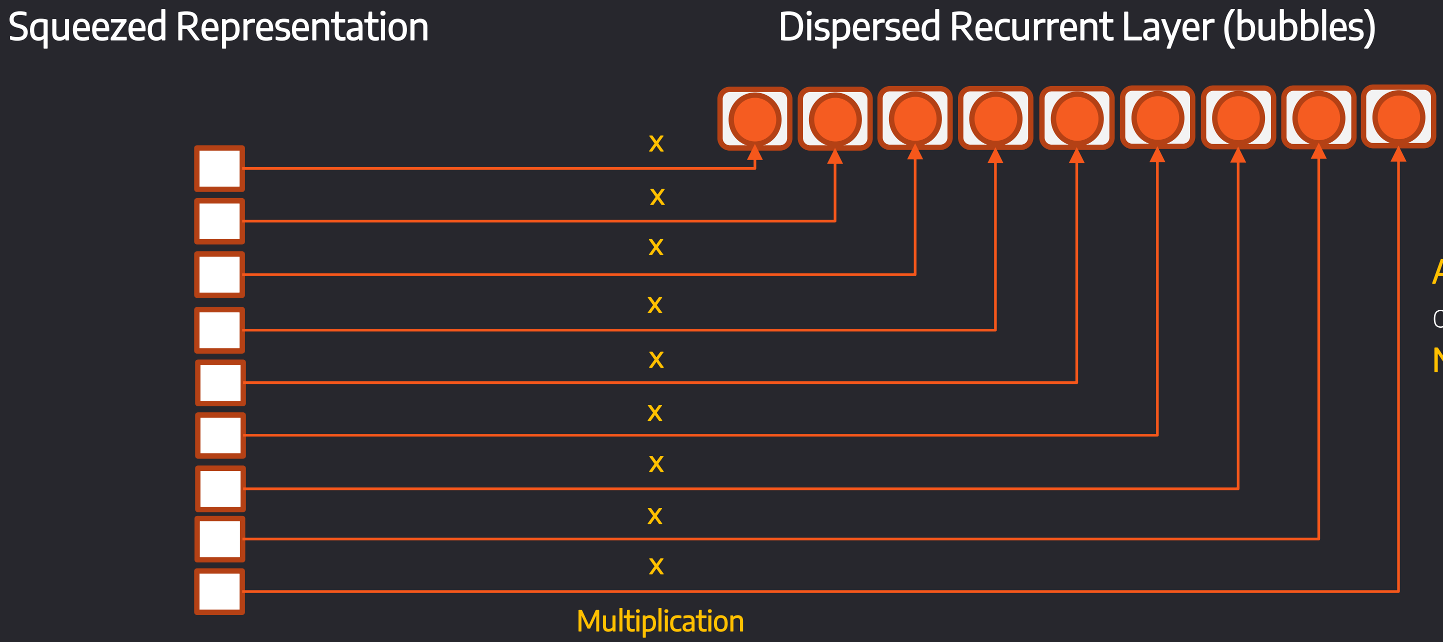
Normalization, Dense and Dropout layers

BUBBLENET ARCHITECTURE (FE LAYERS)



- **FE layers** extract **spatial information** from frames of **secondary input** of the **model**
- Information is used to produce a **Squeezed-representation** to **modulate bubbles**

BUBBLE EXCITATION (ATTENTION MECHANISM)



Attention mechanism based on Squeezed and Excitation Networks [8]

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," IEEE CVPR, pp. 7132–7141, 2017.

BUBBLE EXCITATION (ATTENTION MECHANISM)

Recurrent Bubbles (original)

Weight matrix and hidden state

$$y(i) = \sigma_y\,(W_y h_t + b_y)$$

Activation of each bubble

Activation function

Bias

Sum

+

Modulated Bubbles (excited)

Weight matrix and hidden state

$$(\alpha \times (\sigma_y\,(W_y h_t + b_y))$$

Weighting scalar

Activation function

Bias

where:

$$ht = \sigma_y\,(W_x h_x + U_h h_{t-1} + b_h)$$

- **Attention weighted** response from **bubbles** derives from the **sum** of **original bubbles** and **modulated signal**

DATABASES

Human Motion Database (HMDB-51) [9]

6,849 videos
51 different activity classes

UCF-101 Dataset [11]

11,320 videos
101 different activity classes

YUP++ Dynamic Scenes [10]

1,200 videos
20 different scene classes

[9] H. Kuhne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “Hmdb: A large video database for human motion recognition,” in IEEE ICCV, 2011.
[10] C. Feichtenhofer, A. Pinz, and R. Wildes, “Temporal residual networks for dynamic scene recognition,” in IEEE CVPR, 07 2017, pp. 7435–7444
[11] K. Soomro, A. Zamir, and M. Shah, “Ucf101: A dataset of 101 human actions classes from videos in the wild,” CoRR, 2012.

Hyperparameter adjustment

Table 1 – BubbleNET Hyperparameters for each dataset

	HMDB-51	UCF-101	YUP++
Number of Bubbles	256	256	36
Size of Bubbles (units)	24	28	12
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)		
Learning Rate	0.00005		

- BubbleNET major hyperparameters for each evaluated dataset. Type of recurrent layer was also determined through validation tests

EVALUATION OF BUBBLENET

Table 2 – BubbleNET results in comparison to literature approaches

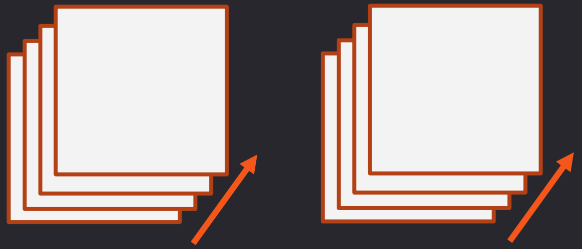
	HDMB-51	UCF-101	YUP++
HAF + BOW/FV	82.48%	-	**92.60%**
ADL + I3D	81.50%	-	91.70%
EvaNET	82.30%	-	-
Two-Stream I3D	80.20%	**97.90%**	-
MARS + RGB + Flow	81.30%	97.80%	-
BubbleNET + I3D RGB + I3D Flow + RNN	76.20%	92.43%	86.75%
BubbleNET + Appearance FE + I3D Flow + RNN	80.20%	95.63%	90.75%
BubbleNET + Excited Bubbles (no sum)	78.70%	85.82%	84.58%
BubbleNET + RNN	82.58%	97.62%	91.70%
BubbleNET + LSTM	**82.60%**	97.20%	91.56%

- Average Recognition **accuracy** of **BubbleNET** and **state-of-the-art** methods

EVALUATION OF BUBBLENET

Table 3 – Variations on BubbleNET recurrent layer

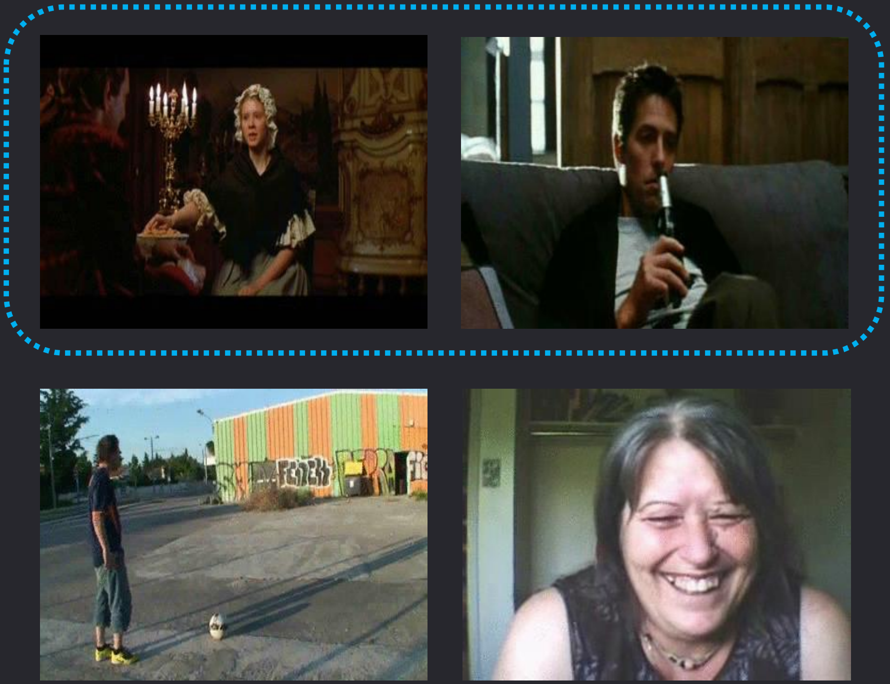
	YUP++
Disperse Recurrent Layer (bubbles)	91.70%
Single Recurrent Layer	81.11%
No Recurrent Layers	88.70%



- Average Recognition **accuracy** for **alternatives** for the **disperse recurrent layer**

CONCLUDING REMARKS AND DISCUSSION

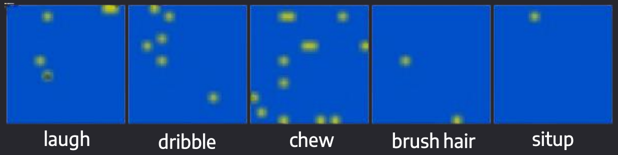
Mean Activation



- **Mean activation** of each bubble used to produce a **signature** for the video classes
- **Similar activities** present **closer** signatures **than dissimilar activities**

CONCLUDING REMARKS AND DISCUSSION

Standard Deviation



- Low standard deviation for video instances of a same class
- BubbleNET captures what characterizes each class

CONCLUDING REMARKS AND DISCUSSION

BubbleNET Results

are comparable to state-of-the-art methods, emerging as the best reported approach for HMDB-51

Class Signatures

support the idea that the disperse bubble layer modularized correlated information, acting as a bag-of-activations

Application in different fields

is a next step of BubbleNET, with tests being considered to be performed in sign-language modeling and gesture recognition

THANK YOU !

Igor L. O. Bastos

igorcrexito@gmail.com, igorlobastos@dcc.ufmg.br

https://github.com/igorcrexito/BubbleNetwork

https://www.researchgate.net/profile/Igor_Bastos2

https://www.linkedin.com/in/igor-bastos-3442481b/

FAPEMIG CAPES CNPq DCC DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO UFMG