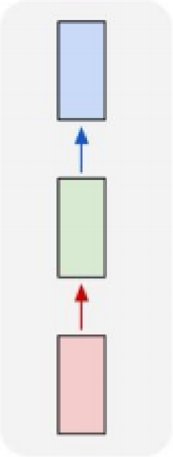
01

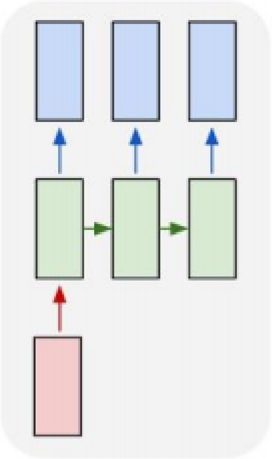
Deep Learning

Deep Learning — Input / Output

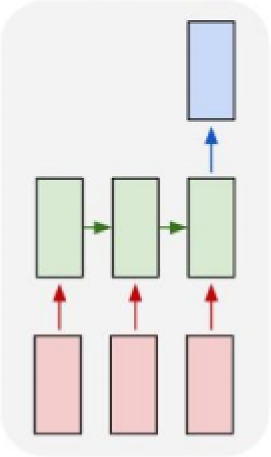
one to one



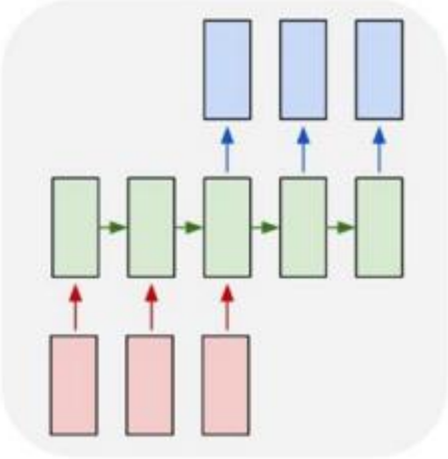
one to sequence



sequence to one

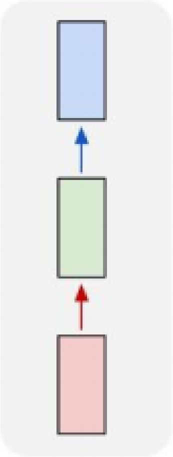


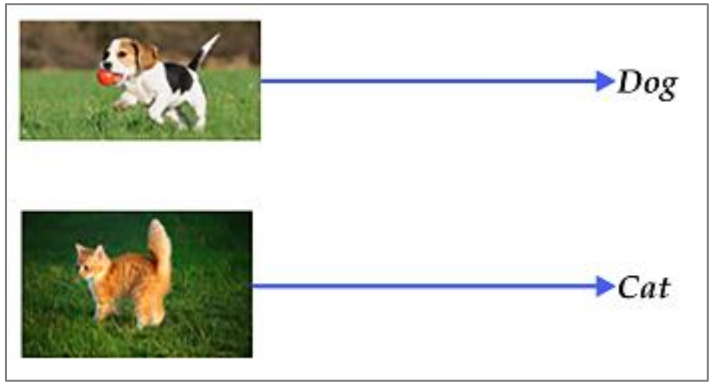
sequence to sequence



Deep Learning — Input / Output

one to one



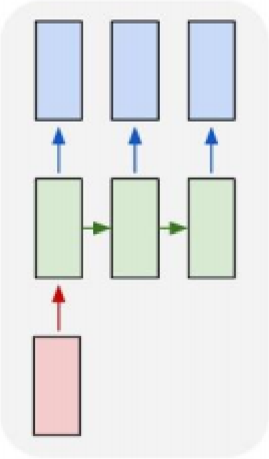


图像分类

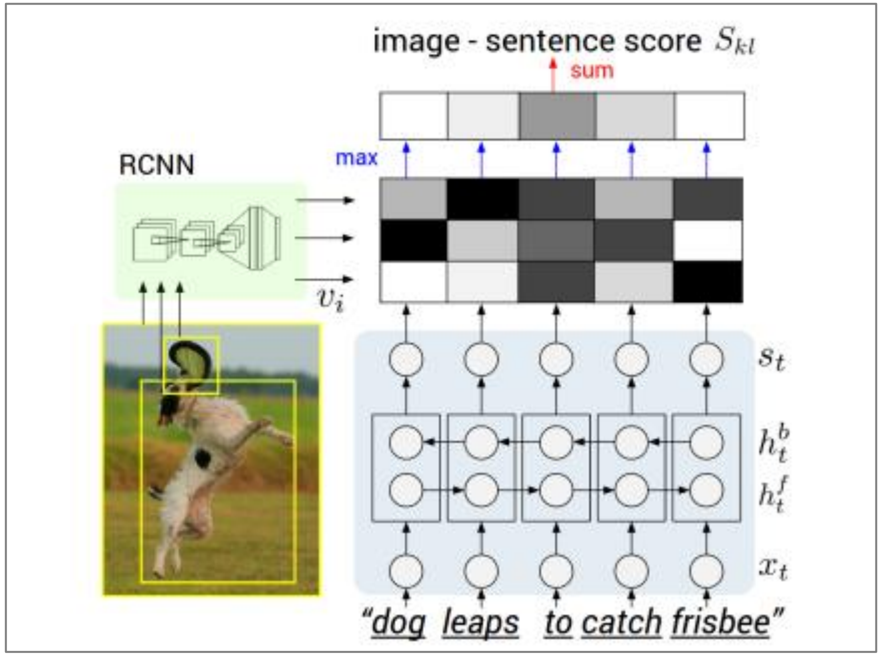
AlexNet[1], VGGNet[2], GoogLeNet[3], ResNet[4]

Deep Learning — Input / Output

one to sequence



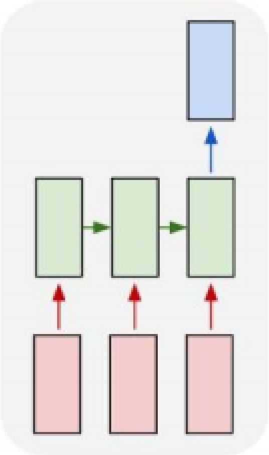
图像描述（字幕）



[5]

Deep Learning — Input / Output

sequence to one

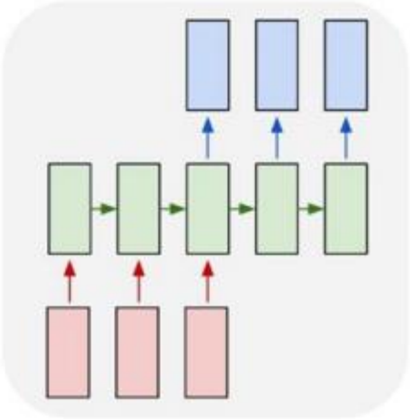


情感分析

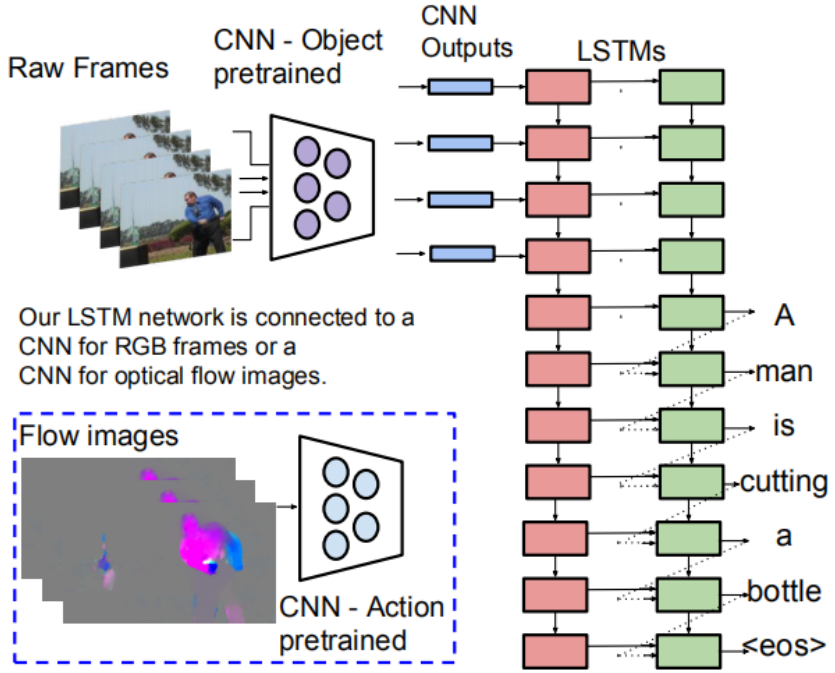
“This” “is” “the” “best” “movie” “ever” → positive

Deep Learning — Input / Output

sequence to sequence



视频描述

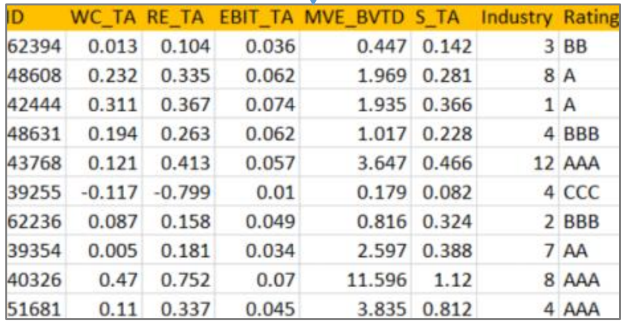


数据集类型

数值类型

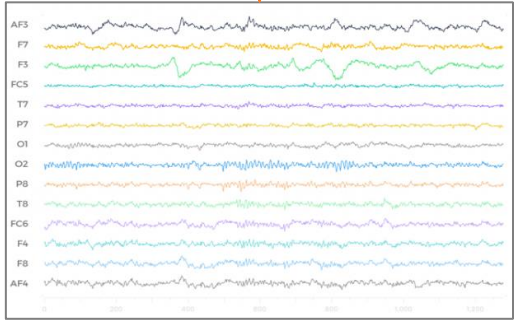
时间序列、
文本数据

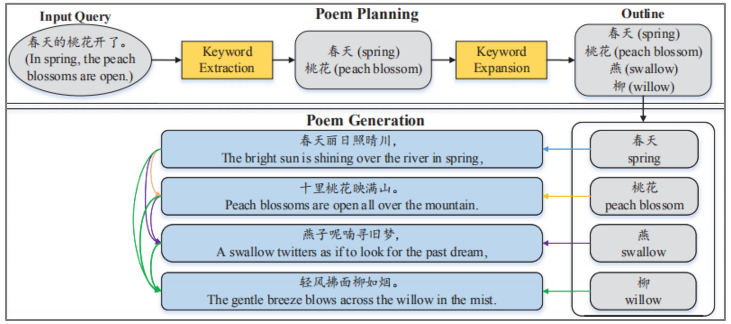
图像数据



ID	WC_TA	RE_TA	EBIT_TA	MVE_BVTD	S_TA	Industry	Rating
62394	0.013	0.104	0.036	0.447	0.142	3	BB
48608	0.232	0.335	0.062	1.969	0.281	8	A
42444	0.311	0.367	0.074	1.935	0.366	1	A
48631	0.194	0.263	0.062	1.017	0.228	4	BBB
43768	0.121	0.413	0.057	3.647	0.466	12	AAA
39255	-0.117	-0.799	0.01	0.179	0.082	4	CCC
62236	0.087	0.158	0.049	0.816	0.324	2	BBB
39354	0.005	0.181	0.034	2.597	0.388	7	AA
40326	0.47	0.752	0.07	11.596	1.12	8	AAA
51681	0.11	0.337	0.045	3.835	0.812	4	AAA

ML, LSTM





Input Query

春天的桃花开了。
(In spring, the peach blossoms are open.)

Poem Planning

Keyword Extraction

春天 (spring)
桃花 (peach blossom)

Keyword Expansion

Outline

春天 (spring)
桃花 (peach blossom)
燕 (swallow)
柳 (willow)

Poem Generation

春天丽日照晴川，
The bright sun is shining over the river in spring,

十里桃花映满山。
Peach blossoms are open all over the mountain.

燕子呢喃寻旧梦，
A swallow twitters as if to look for the past dream,

轻风拂面柳如烟。
The gentle breeze blows across the willow in the mist.

春天
spring

桃花
peach blossom

燕
swallow

柳
willow

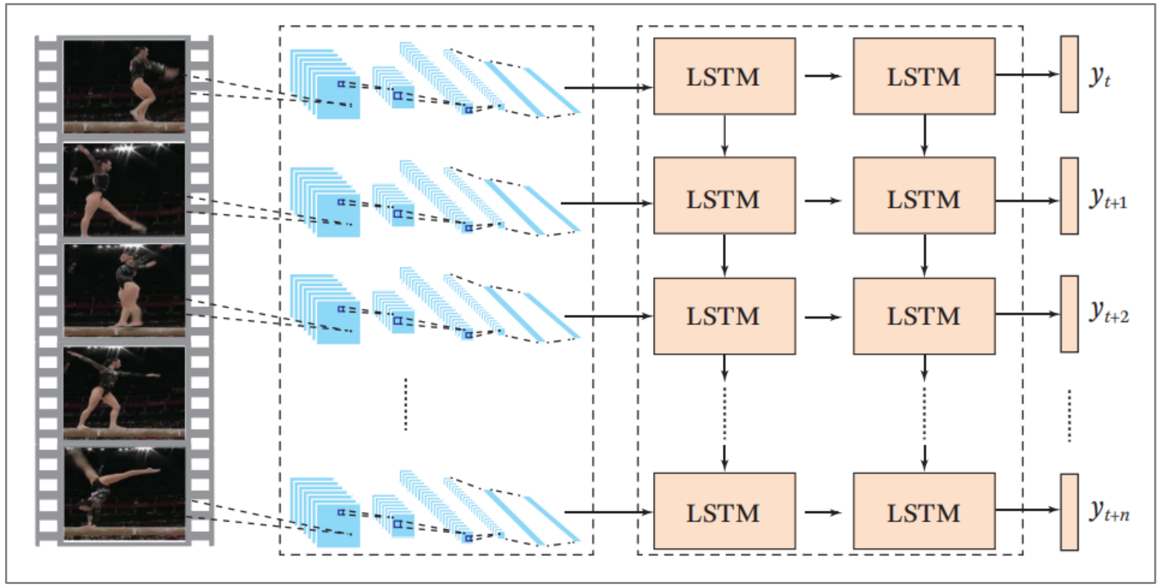
LSTM

CNN

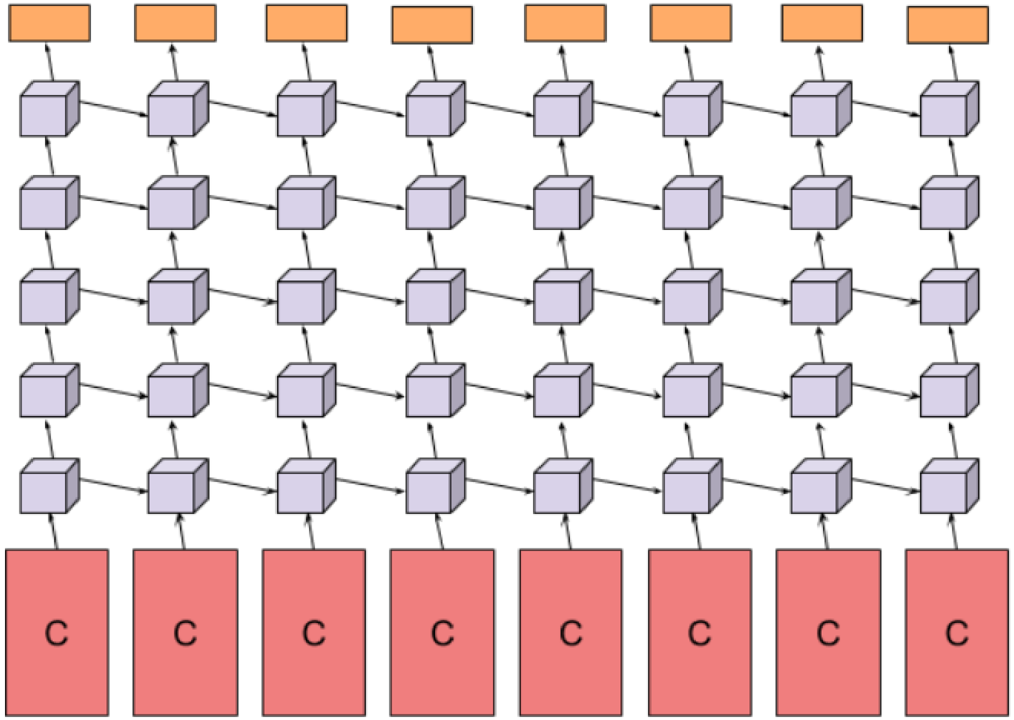
02

Deep Learning for Video Application

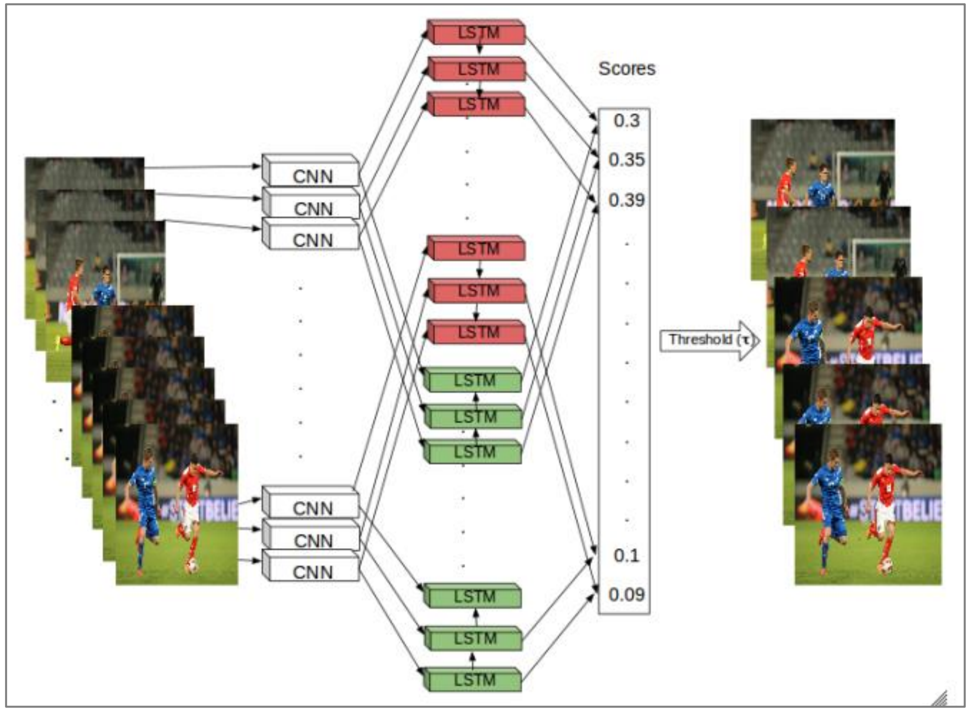
Video Classification



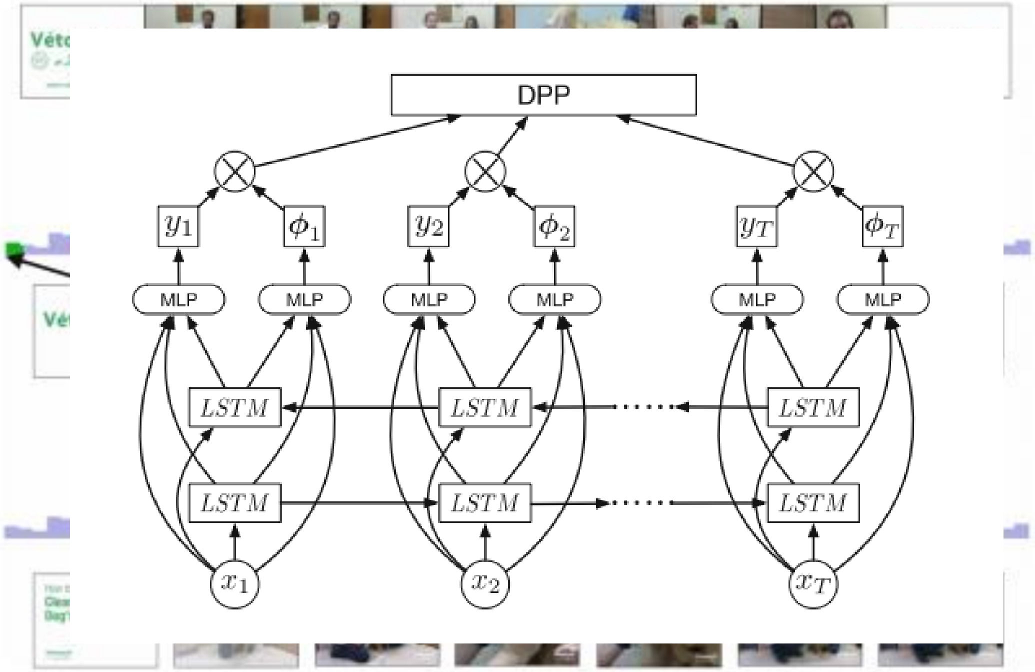
[6]



Video Summarization

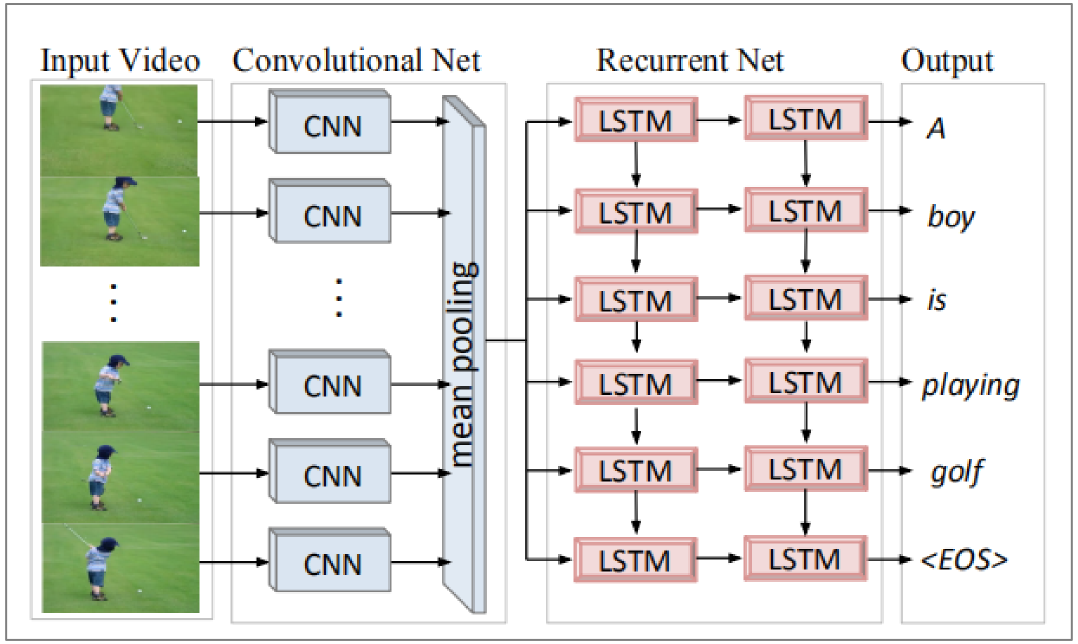


[8]

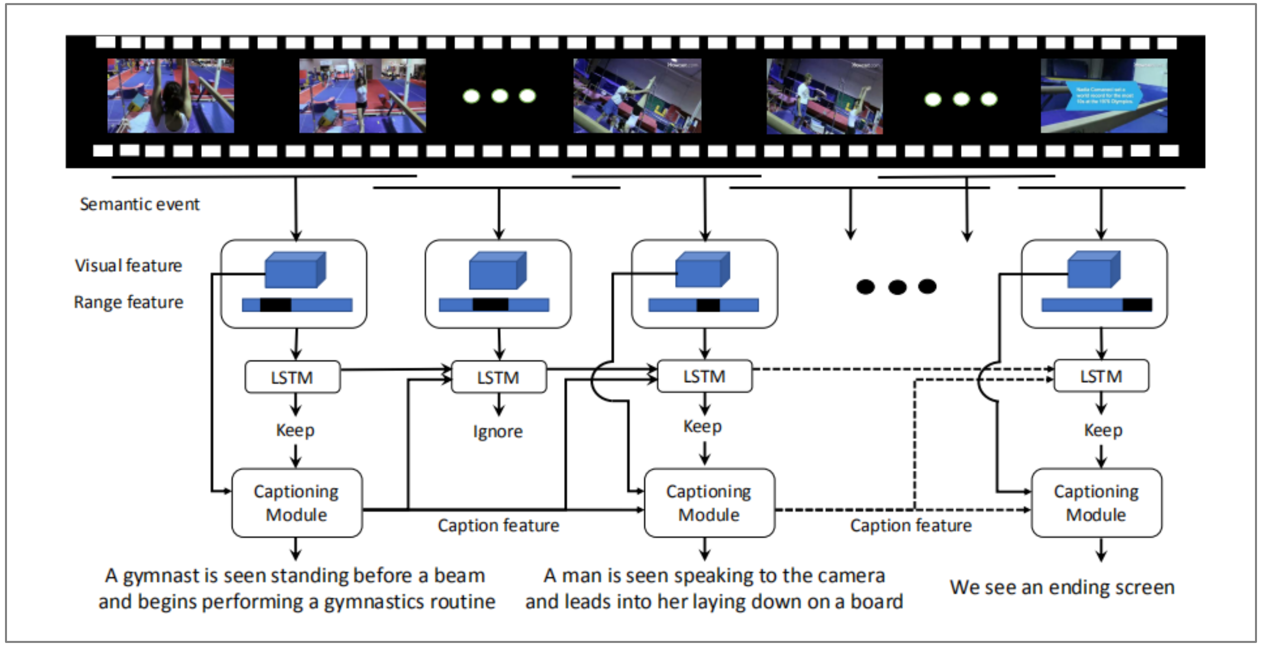


[9]

Video Captioning



[10]


Semantic event
Visual feature
Range feature
LSTM
Keep
Ignore
Keep
Keep
Captioning Module
Caption feature
Caption feature
A gymnast is seen standing before a beam and begins performing a gymnastics routine
A man is seen speaking to the camera and leads into her laying down on a board
We see an ending screen

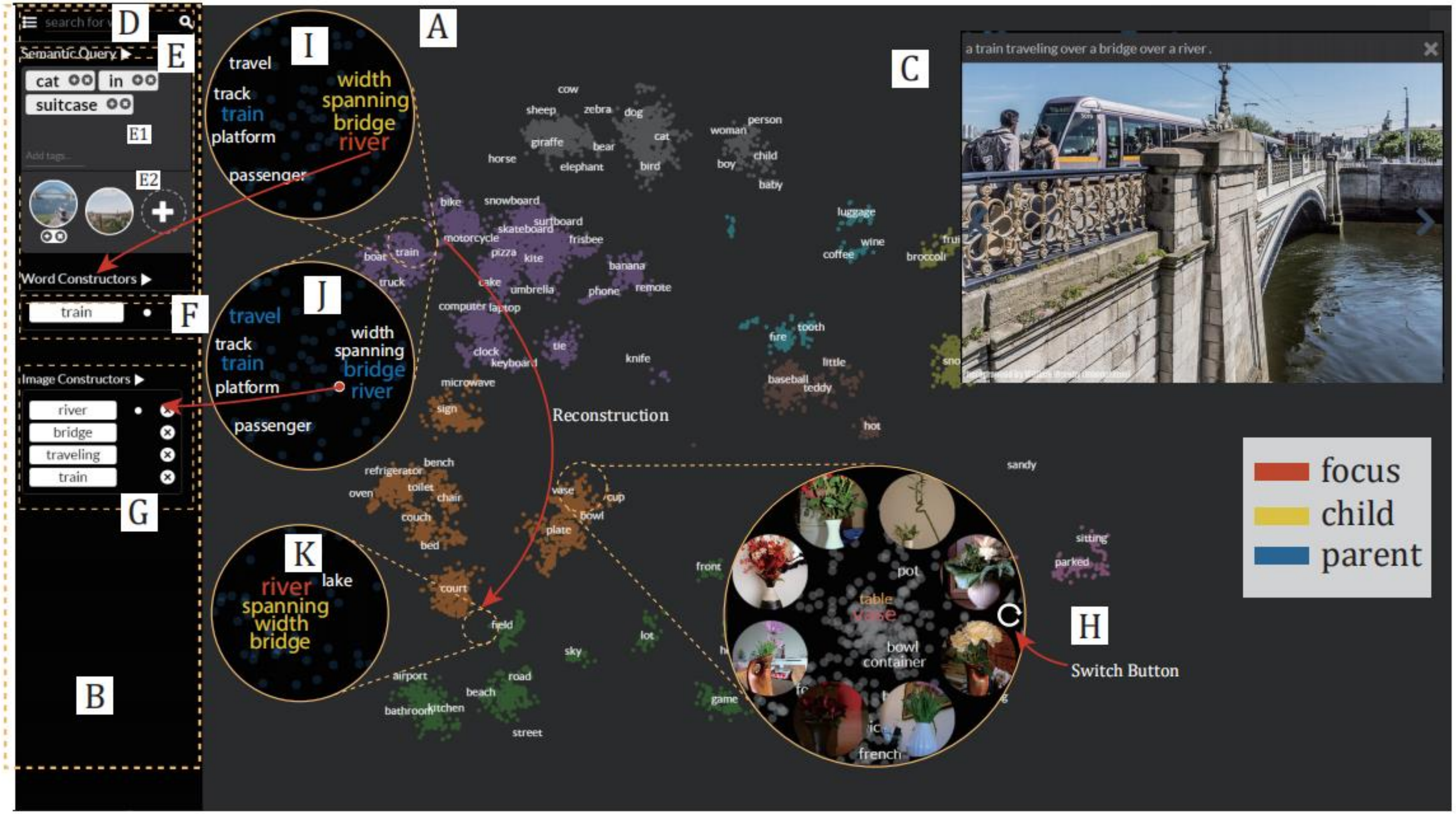
[11]

A Semantic-based Method for Visualizing Large Image Collections

Xiao Xie, Xiwen Cai, Junpei Zhou, Nan Cao, Yingcai Wu

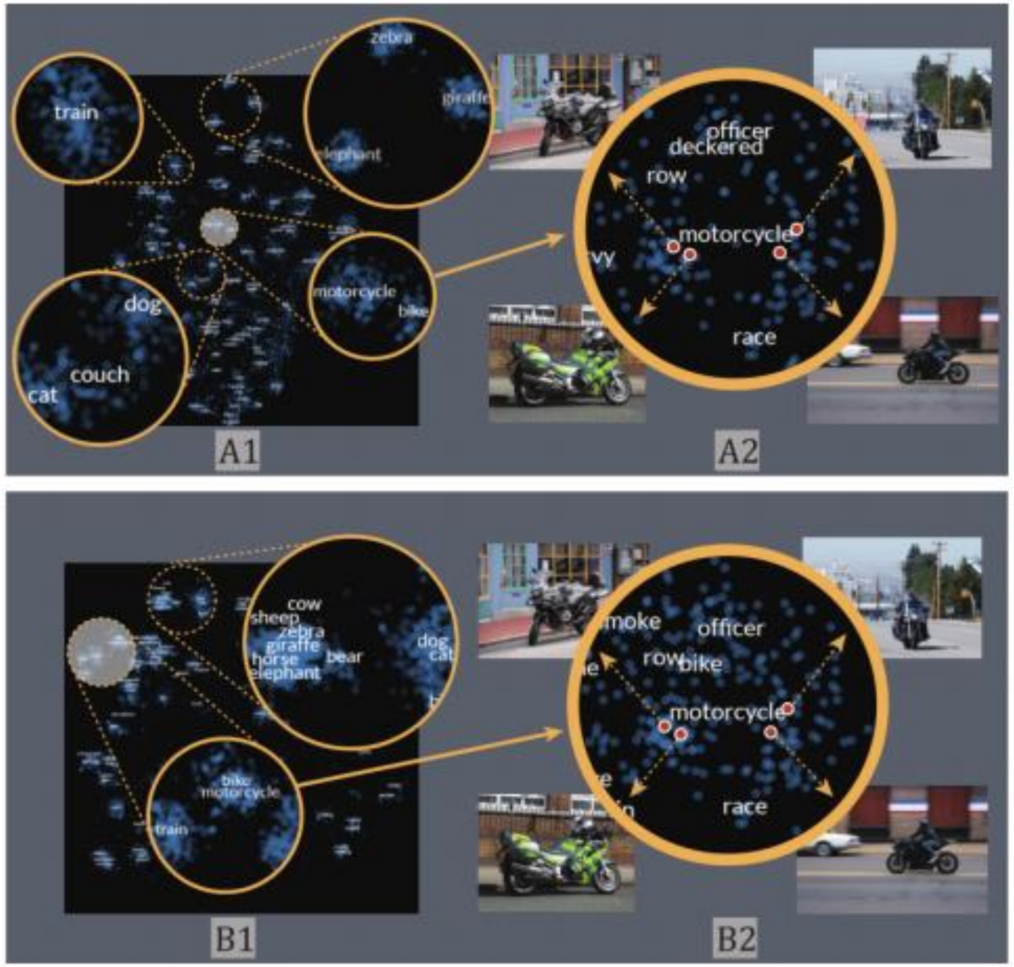
IEEE Transactions on Visualization and Computer Graphics (2018)

Interface

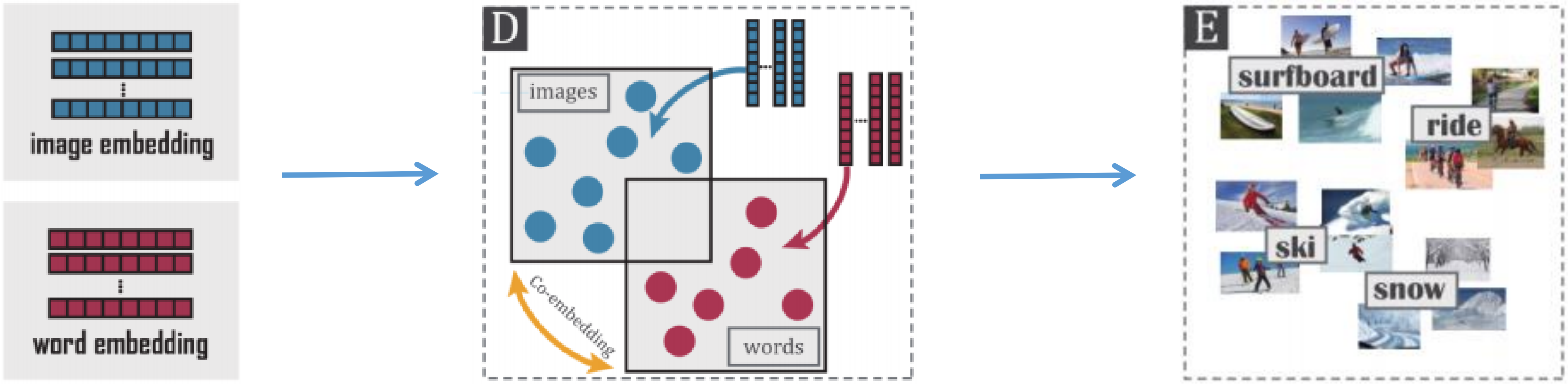


a train traveling over a bridge over a river .

Previous Methods VS Co-embedding



Co-embedding



Semantic Information Extract

The man at bat readies to swing at the pitch while the umpire looks on.

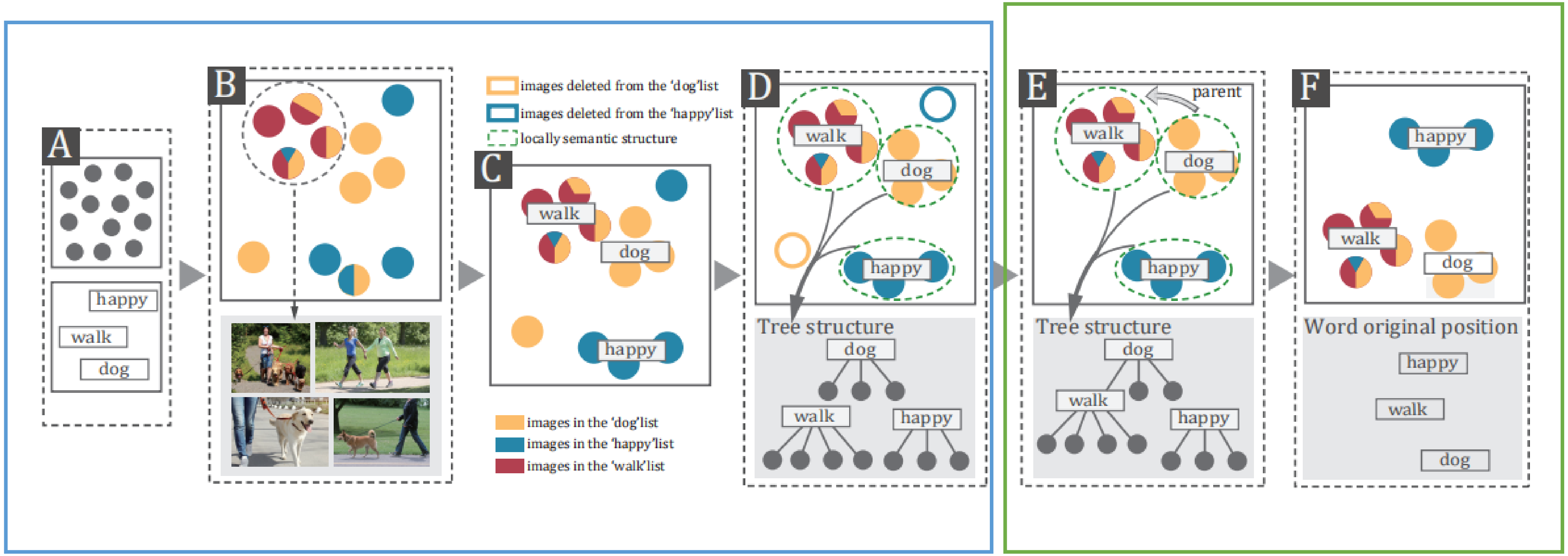
A large bus sitting next to a very tall building.

A horse carrying a large load of hay and two people sitting on it.

Bunk bed with a narrow shelf sitting underneath it.

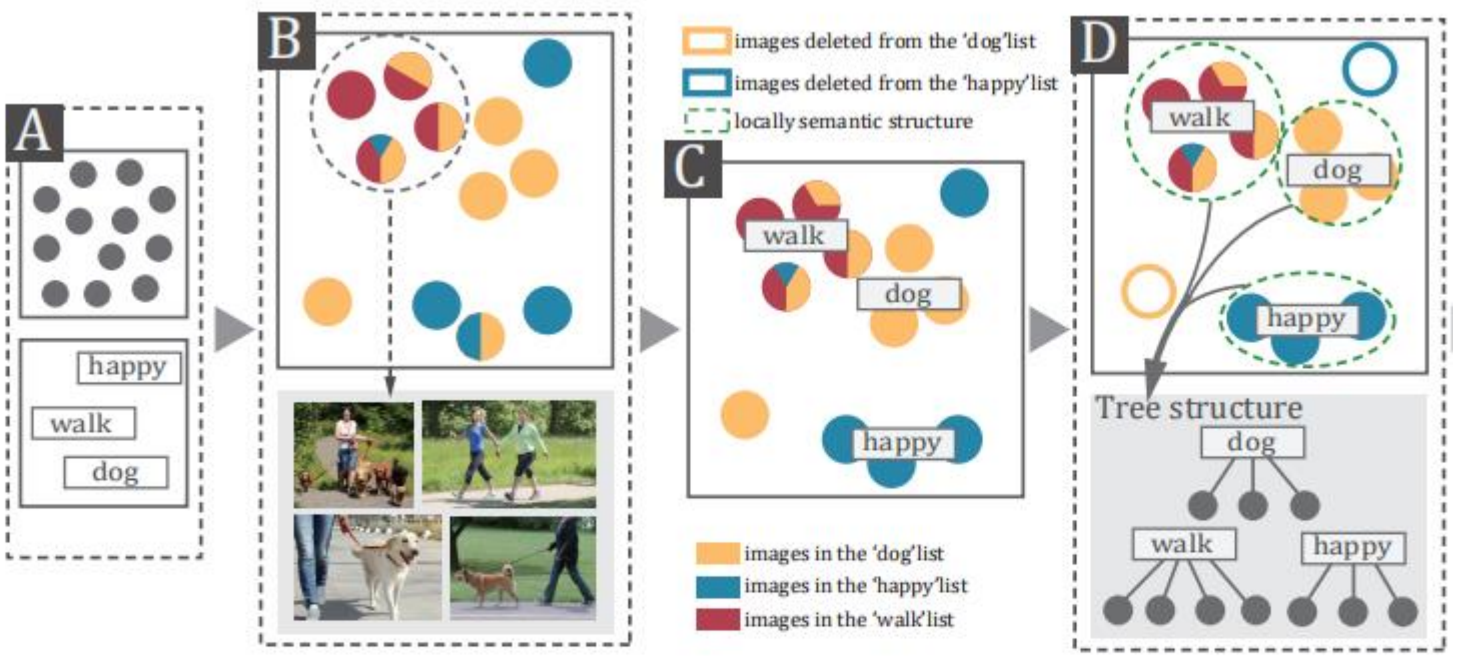
MSCOCO Dataset (80000+ * 5)

Co-embedding



Co-embedding

- Obtaining Local Semantic Structures



A
happy
walk
dog
B
images deleted from the 'dog'list
images deleted from the 'happy'list
locally semantic structure
C
walk
dog
happy
images in the 'dog'list
images in the 'happy'list
images in the 'walk'list
D
walk
dog
happy
Tree structure
dog
walk
happy

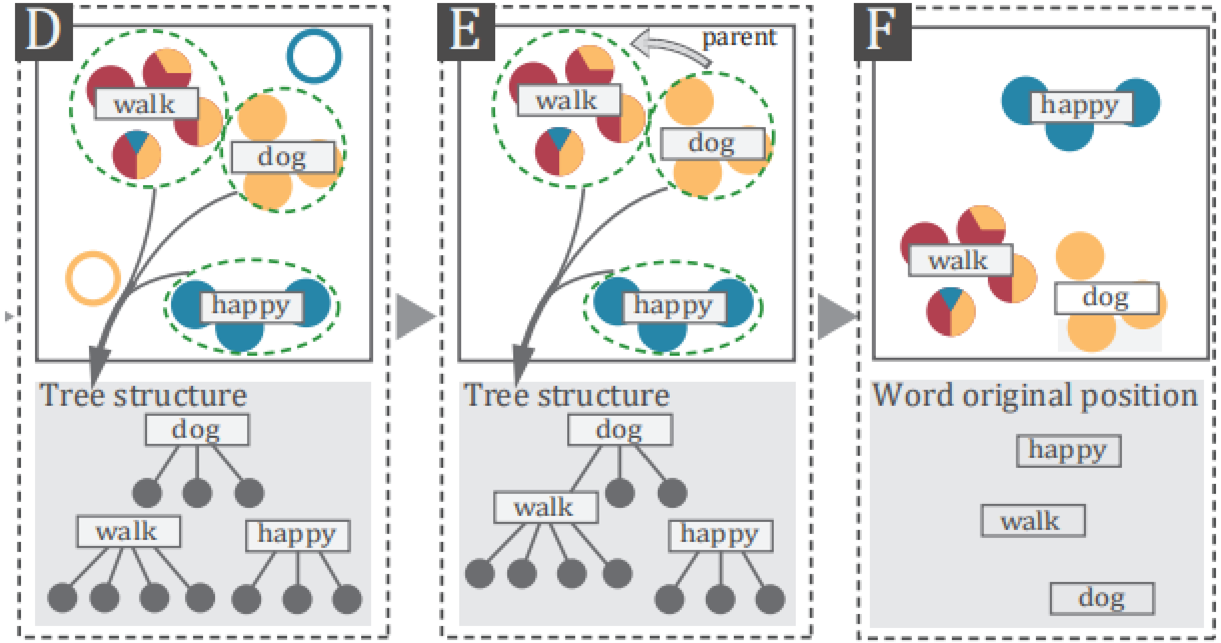
$$Simi(W_i, I_j) = 1 - \min_{W_k \in C_j} d(W_i, W_k)$$

$$\mathcal{I}_{W_i} = \{I_j \mid I_j \in \mathcal{I},\ Simi(W_i, I_j) \geq MinSimi\}$$

$$\mathcal{W}_{I_j} = \{W_i \mid W_i \in \mathcal{W},\ I_j \in \mathcal{I}_{W_i}\}$$

Co-embedding

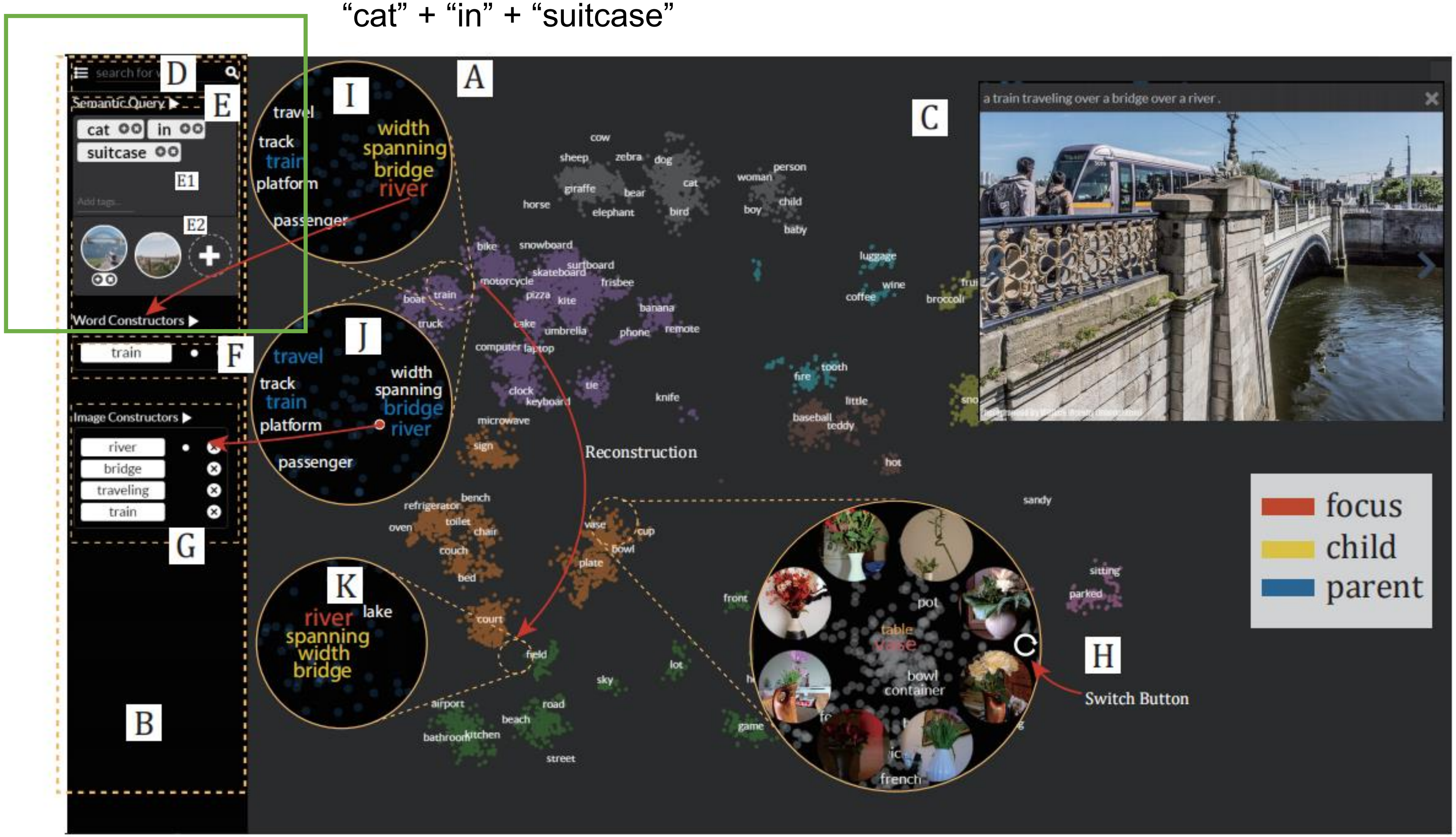
- Reconstruct Images in Semantic Space



$$Freq(W_i, W_j) = Freq(W_j, W_i) = |\mathcal{I}_{W_i} \cap \mathcal{I}_{W_j}|$$

$$\mathbf{CF}_{ij} = \frac{Freq(W_i, W_j)}{Freq(W_i)}$$

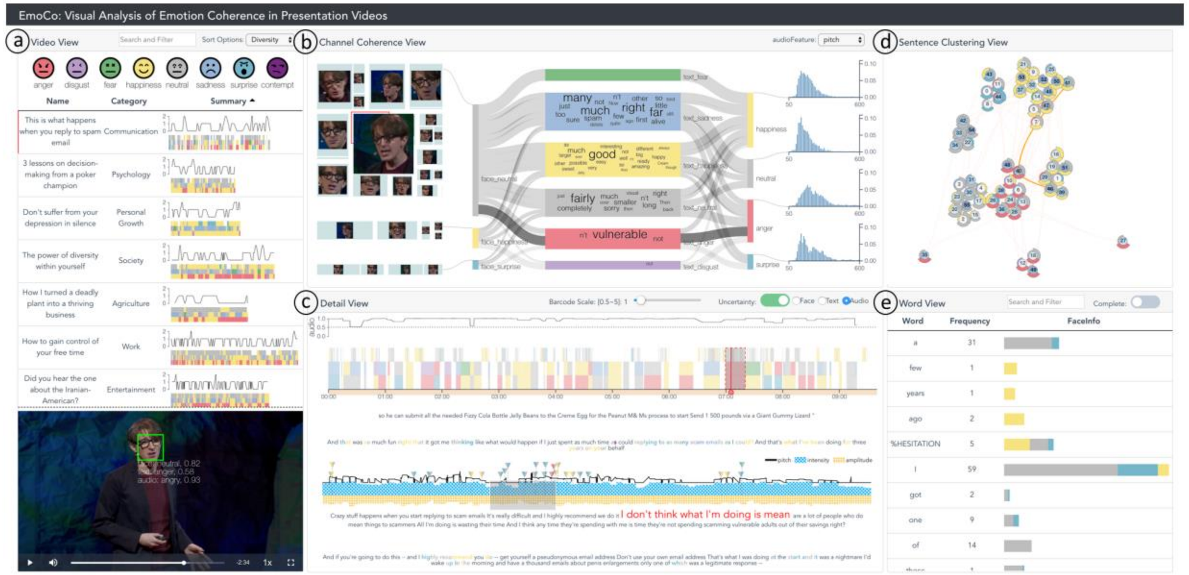
“dog” + “cat” + “image of the beach”
“cat” + “in” + “suitcase”



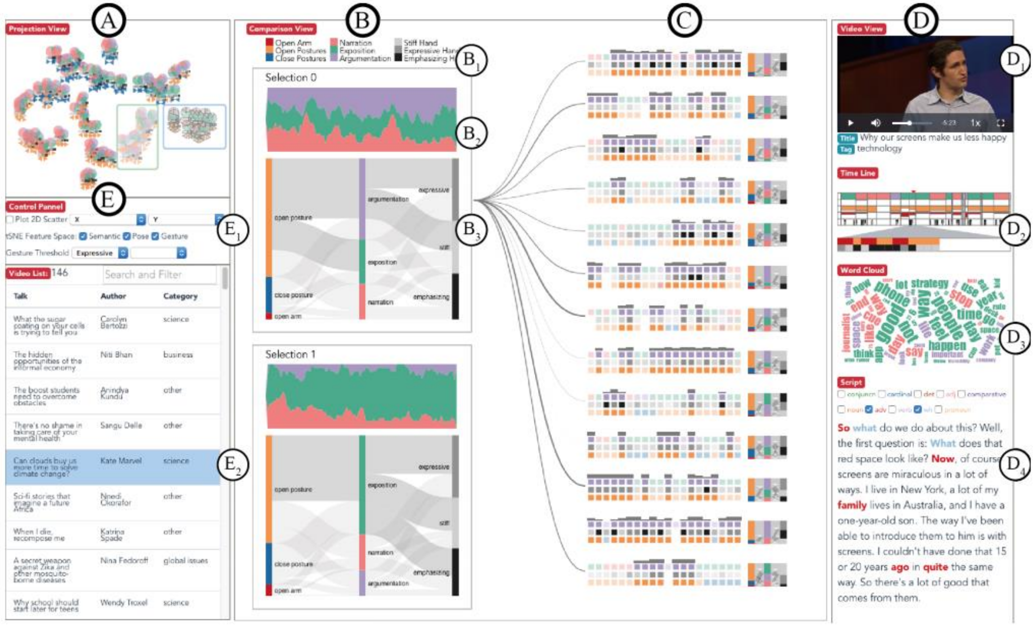
EmotionCues: Emotion-Oriented Visual Summarization of Classroom Videos

Haipeng Zeng, Xinhuan Shu, Yanbang Wang, Yong Wang,
Liguo Zhang,Ting-Chuen Pong, and Huamin Qu

IEEE Transactions on Visualization and Computer Graphics (2020)

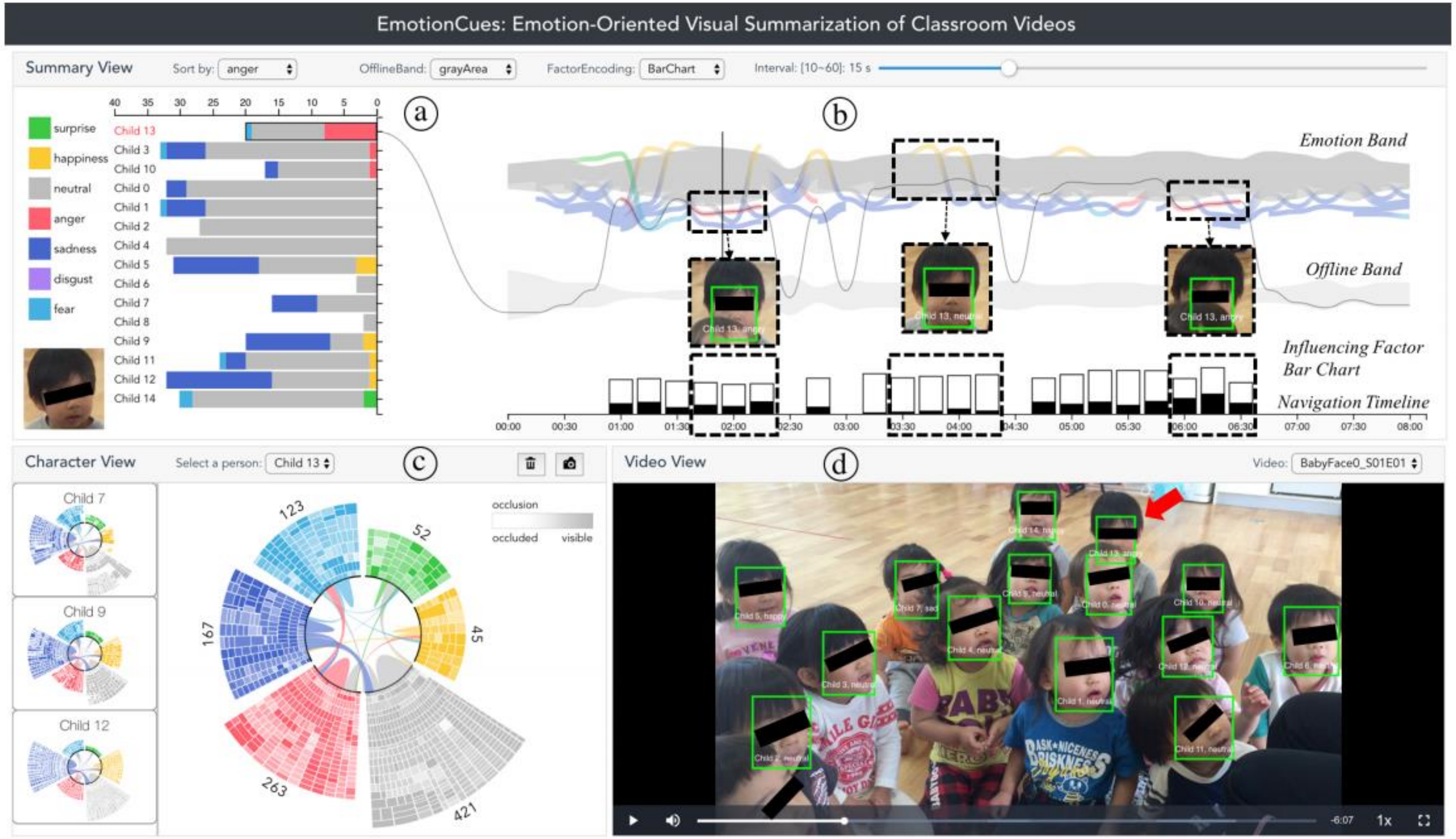

EmoCo: Visual Analysis of Emotion Coherence in Presentation Videos
a Video View
b Channel Coherence View
c Detail View
d Sentence Clustering View
e Word View
anger
disgust
fear
happiness
neutral
sadness
surprise
contempt
Name
Category
Summary
Word
Frequency
FaceInfo

[12]


A
B
C
D
E
Projection View
Comparison View
Video View
Control Pannel
Video List
Selection 0
Selection 1
Time Line
Word Cloud
Script
Open Arm
Open Postures
Close Postures
Narration
Exposition
Argumentation
Stiff Hand
Expressive Hand
Emphasizing Hand
B1
B2
B3
D1
D2
D3
D4
E1
E2
Talk
Author
Category

[13]

Interface



EmotionCues: Emotion-Oriented Visual Summarization of Classroom Videos

Summary View

Sort by: anger

OfflineBand: grayArea

FactorEncoding: BarChart

Interval: [10~60]: 15 s

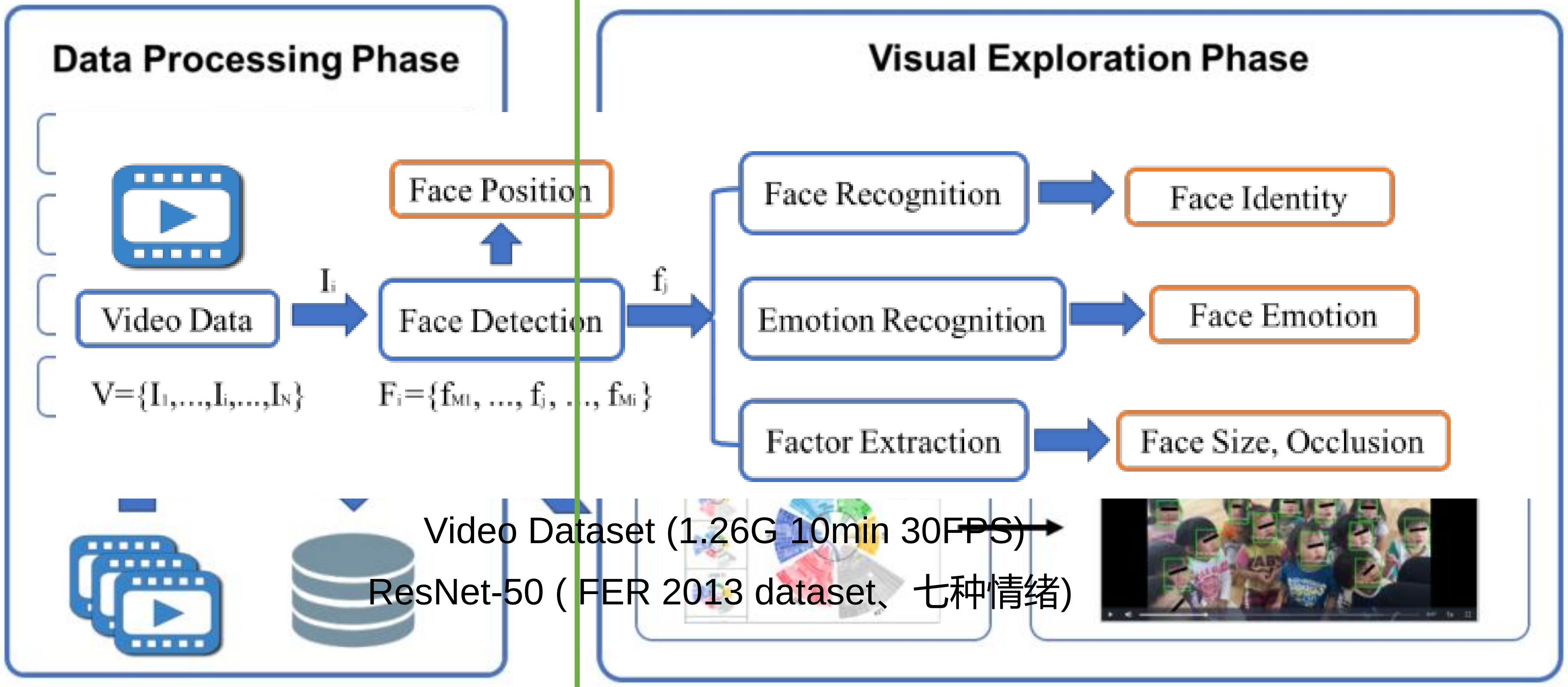
Character View

Select a person: Child 13

Video View

Video: BabyFace0_S01E01

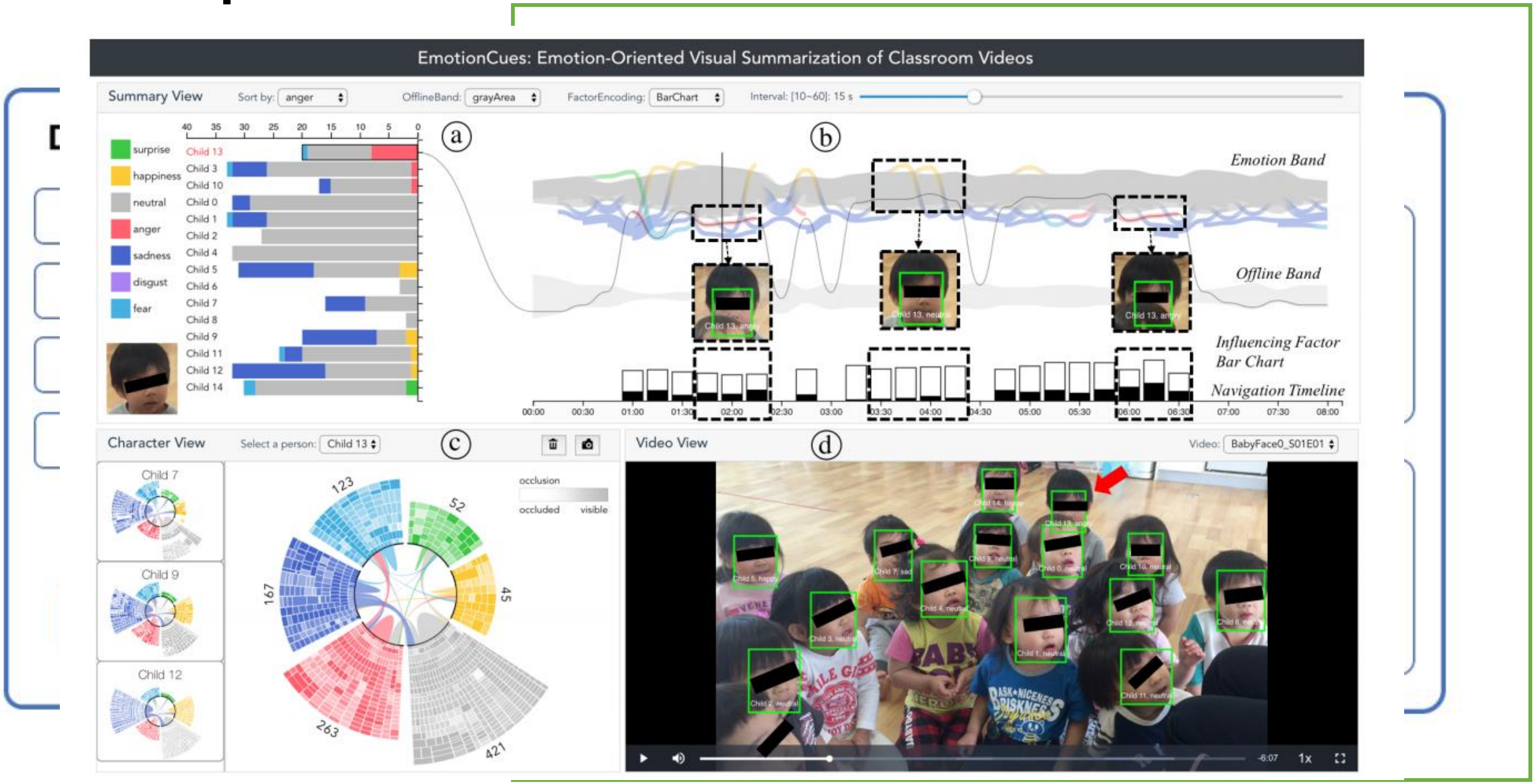
Data Processing Phase



Video Dataset (1.26G 10min 30FPS)

ResNet-50 (FER 2013 dataset、七种情绪)

Visual Exploration Phase



EmotionCues: Emotion-Oriented Visual Summarization of Classroom Videos

Summary View Sort by: anger OfflineBand: grayArea FactorEncoding: BarChart Interval: [10~60]: 15 s

Character View Select a person: Child 13

Video View Video: BabyFace0_S01E01

References

- [1] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.
- [2] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
- [3] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
- [4] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
- [5] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
- [6] Wu, Zuxuan, et al. "Deep learning for video classification and captioning." Frontiers of multimedia research. 2017. 3-29.

- [7] Yue-Hei Ng, Joe, et al. "Beyond short snippets: Deep networks for video classification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
- [8] Naik, Juhi. "DeepVideo: Video Summarization using Temporal Sequence Modelling."
- [9] Zhang, Ke, et al. "Video summarization with long short-term memory." European conference on computer vision. Springer, Cham, 2016.
- [10] Venugopalan, Subhashini, et al. "Translating videos to natural language using deep recurrent neural networks." arXiv preprint arXiv:1412.4729 (2014).
- [11] Xiong, Yilei, Bo Dai, and Dahua Lin. "Move forward and tell: A progressive generator of video descriptions." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
- [12] Zeng, Haipeng, et al. "EmoCo: Visual analysis of emotion coherence in presentation videos." IEEE Transactions on Visualization and Computer Graphics 26.1 (2019): 927-937.
- [13] Wu, Aoyu, and Huamin Qu. "Multimodal analysis of video collections: Visual exploration of presentation techniques in ted talks." IEEE Transactions on Visualization and Computer Graphics (2018).