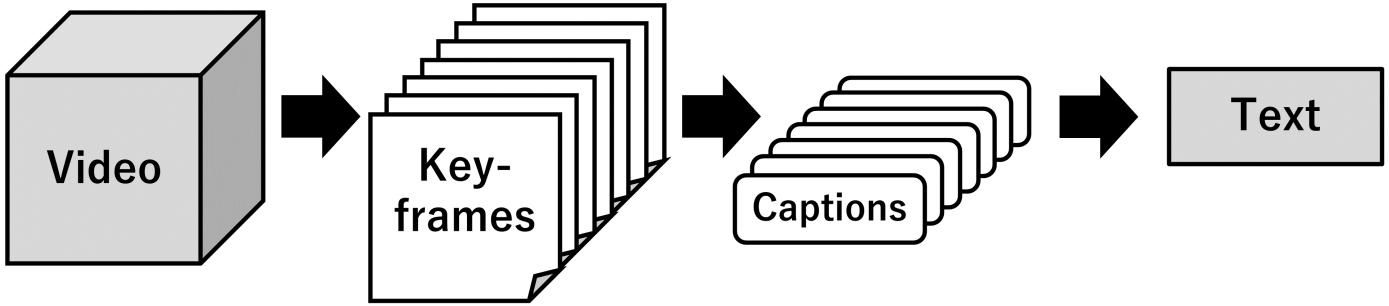
Nagaoka University of Technology at TRECVID 2023: Video to Text

Team kslab, Nagaoka University of Technology
Mutsuki Ishii
Shungo Kubosaka
Takashi Yukawa

Outline

1. Our Previous System
2. Last Year Observations & This Year Targets
3. Frame Extraction Phase
4. Environmental Sound Classification Phase
5. Result
6. Observations
7. Conclution

Our Previous System



- Three phases
  - Frame extraction: GoogLeNet, Kernel Temporal Segmentation
  - Captioning: OFA
  - Aggregation: Lexrank

Last Year Observations & This Year Targets

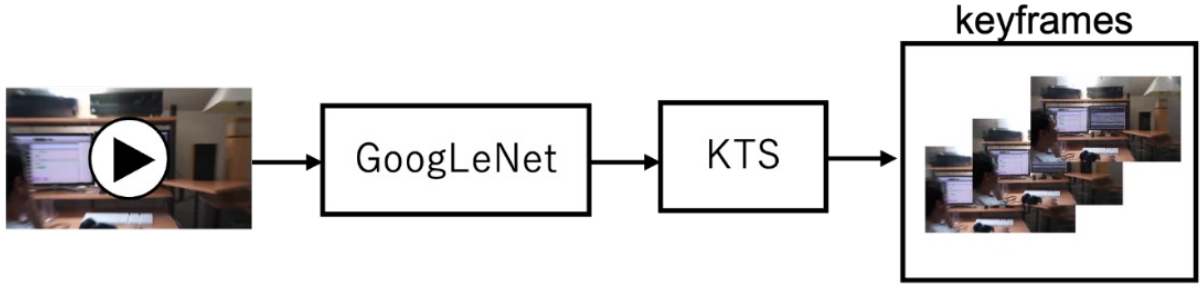
- Captioning using OFA has significantly improved captioning accuracy in last year.

- As a new approach, aiming to incorporate audio in our system.

➡ **Improving the frame extraction method**

➡ **Developing a new phase using audio**

Frame Extraction Phase

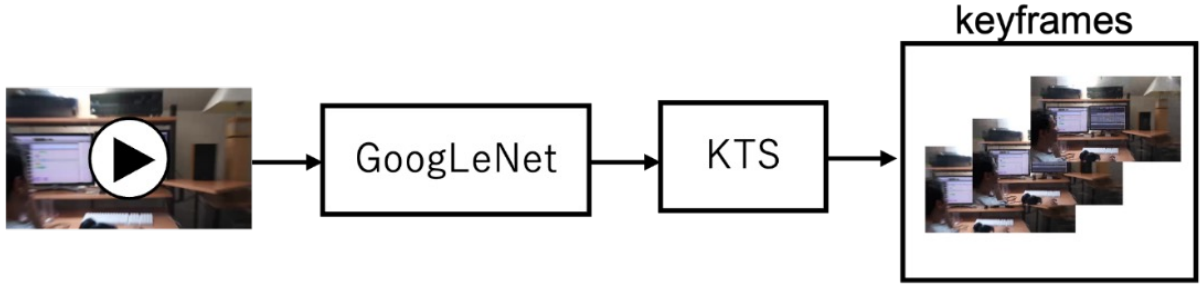


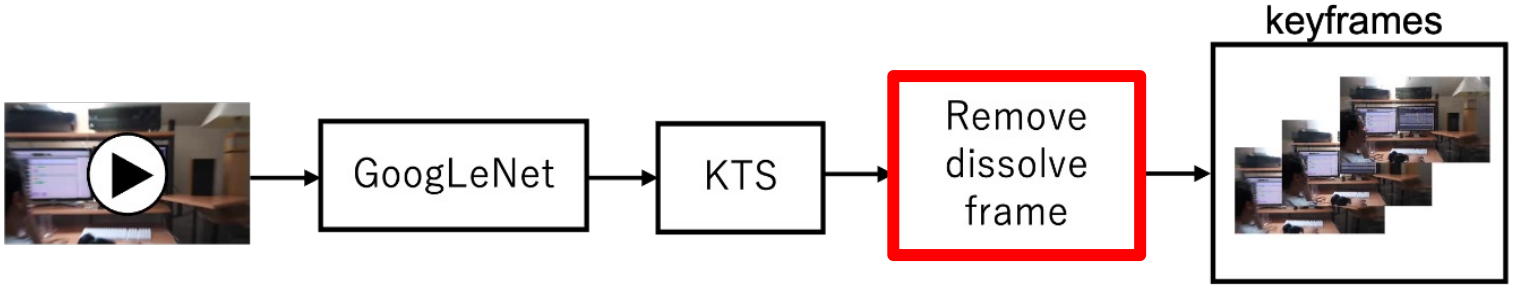
GoogLeNet: Extraction the feature amount for each frame of the video

Kernel Temporal Segmentation: Selecting seven images as keyframes*

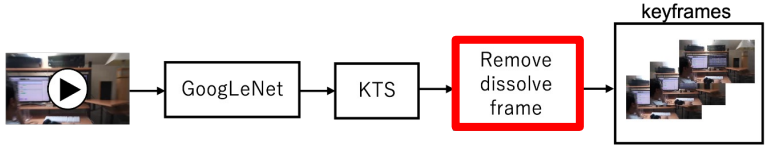
*the sum of five frames with large feature amounts extracted by GoogLeNet and including the first and last frames

Frame Extraction Phase





Frame Extraction Phase



[1]Ioka's
Detection of Dissolve Scene method

Comparing the amount of change in each pixel before and after each gray-scaled frame.

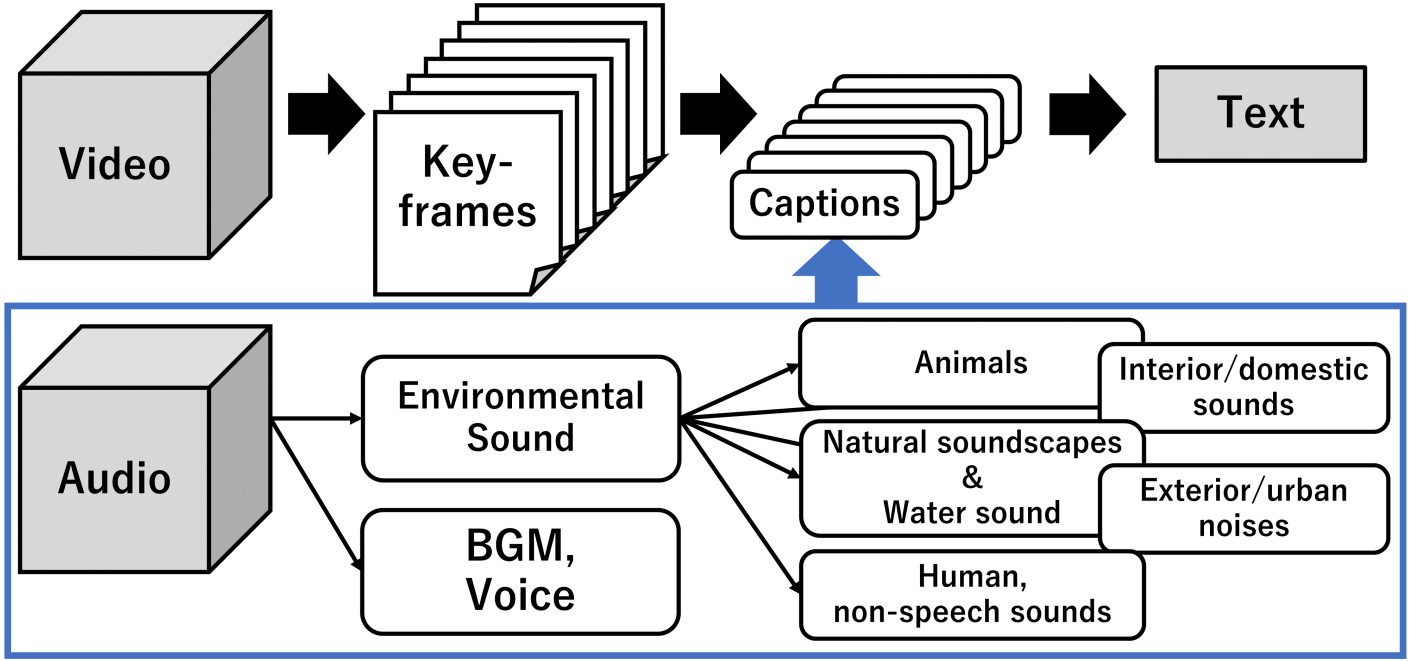
[2]Nagasaka and Tanaka's
Scene-Change Detection method

Calculating the feature by performing a chi-square test on the distribution of RGB values per block between consecutive frames.

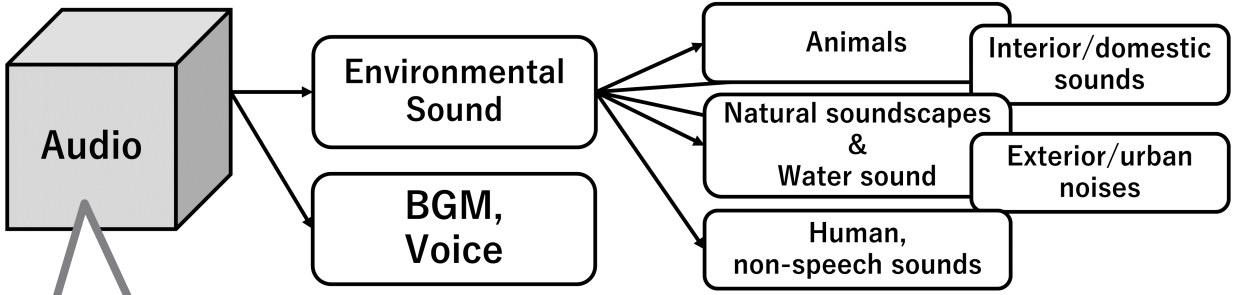
Split video into frames → By ①, the pixel There are many changes Frame extraction → Create a filter from the results of ② and apply it to ①. filter the results → above threshold frame extraction

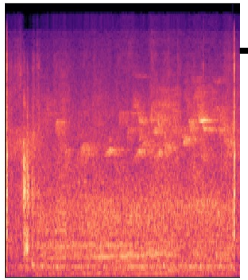
[1] M. Ioka, Detection of dissolve scene change in motion picture, In Proceedings of the 51st National Convention of IPSJ, no.6S-8, pp.247-248, Sept.1995
[2] A. Nagasaka, and Y. Tanaka, Automatic scene-change detection method for video works, In Proceedings of the 40th National Convention of IPSJ, no.1Q-5, pp.642-643, Mar.1990

Environmental Sound Classification Phase



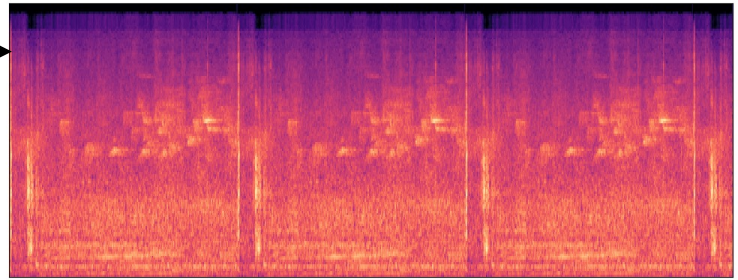
Environmental Sound Classification Phase





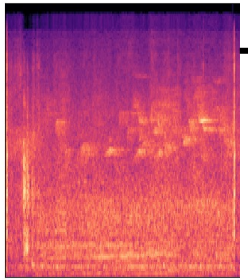
mel spectrogram

The audio data is aligned by looping to make it 16 sec.
An equal amount of sample added white noise is prepared.



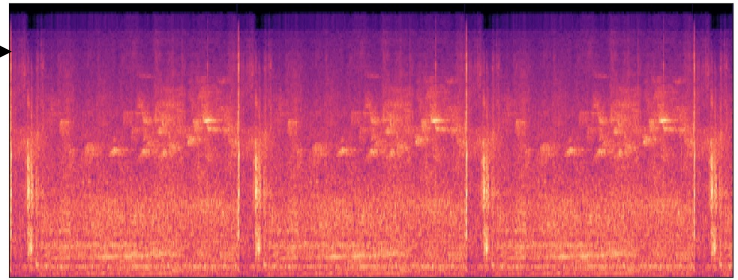
Environmental Sound Classification Phase

Dataset	Type	Time(s)	Sample
ESC-50	Environmental	5	2000
VoxConverse	Human voice	5~15	2240
Free Music Archive	BGM	5~15	2068

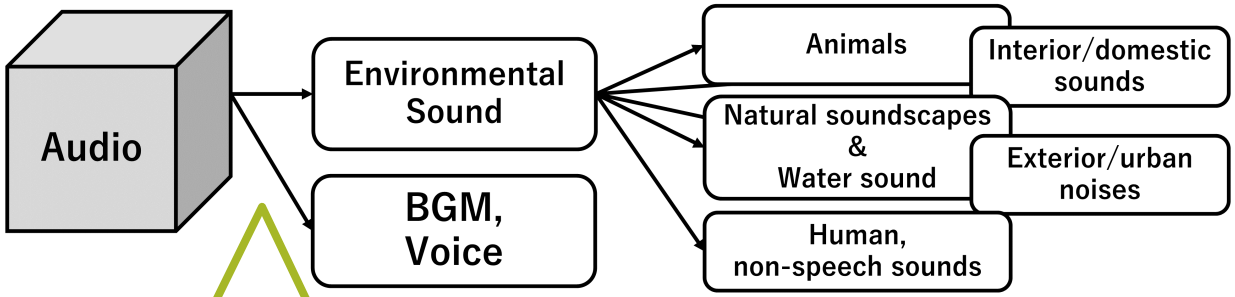


mel spectrogram

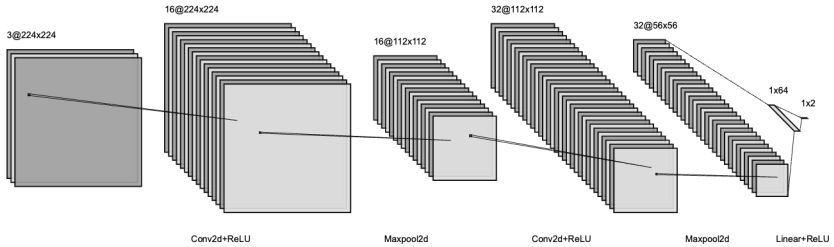
The training data is aligned by looping to make it 16 sec.
An equal amount of sample added white noise is prepared.



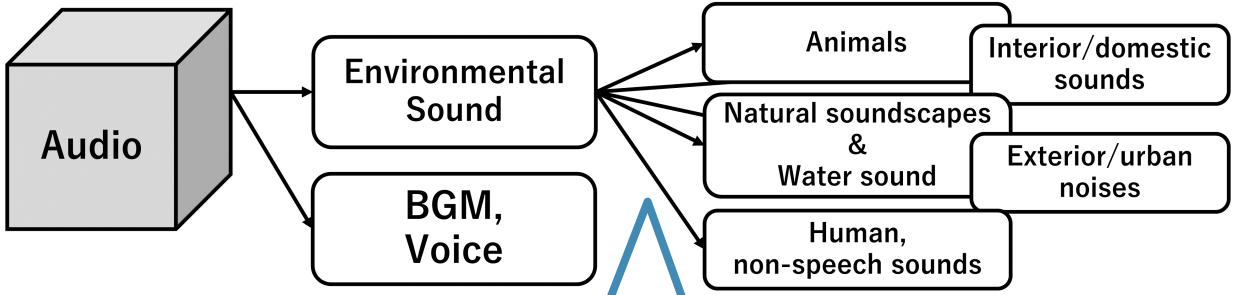
Environmental Sound Classification Phase



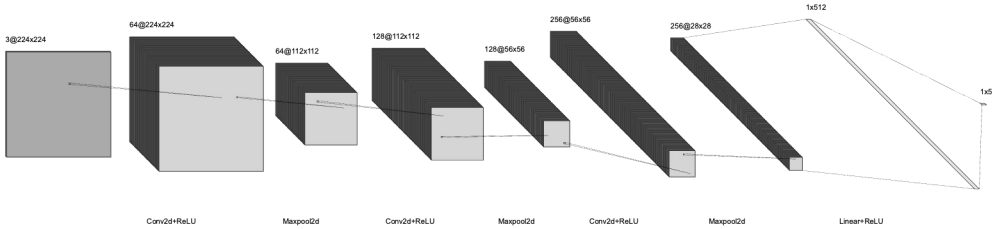
- Train data
  - ESC-50
  - VoxConverse
  - Free Music Archive
- CNN
  - Mel spectrogram



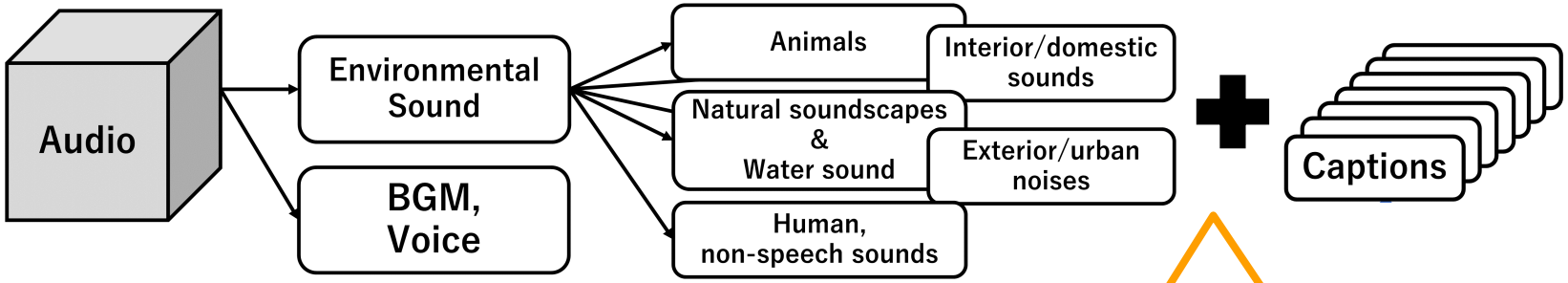
Environmental Sound Classification Phase



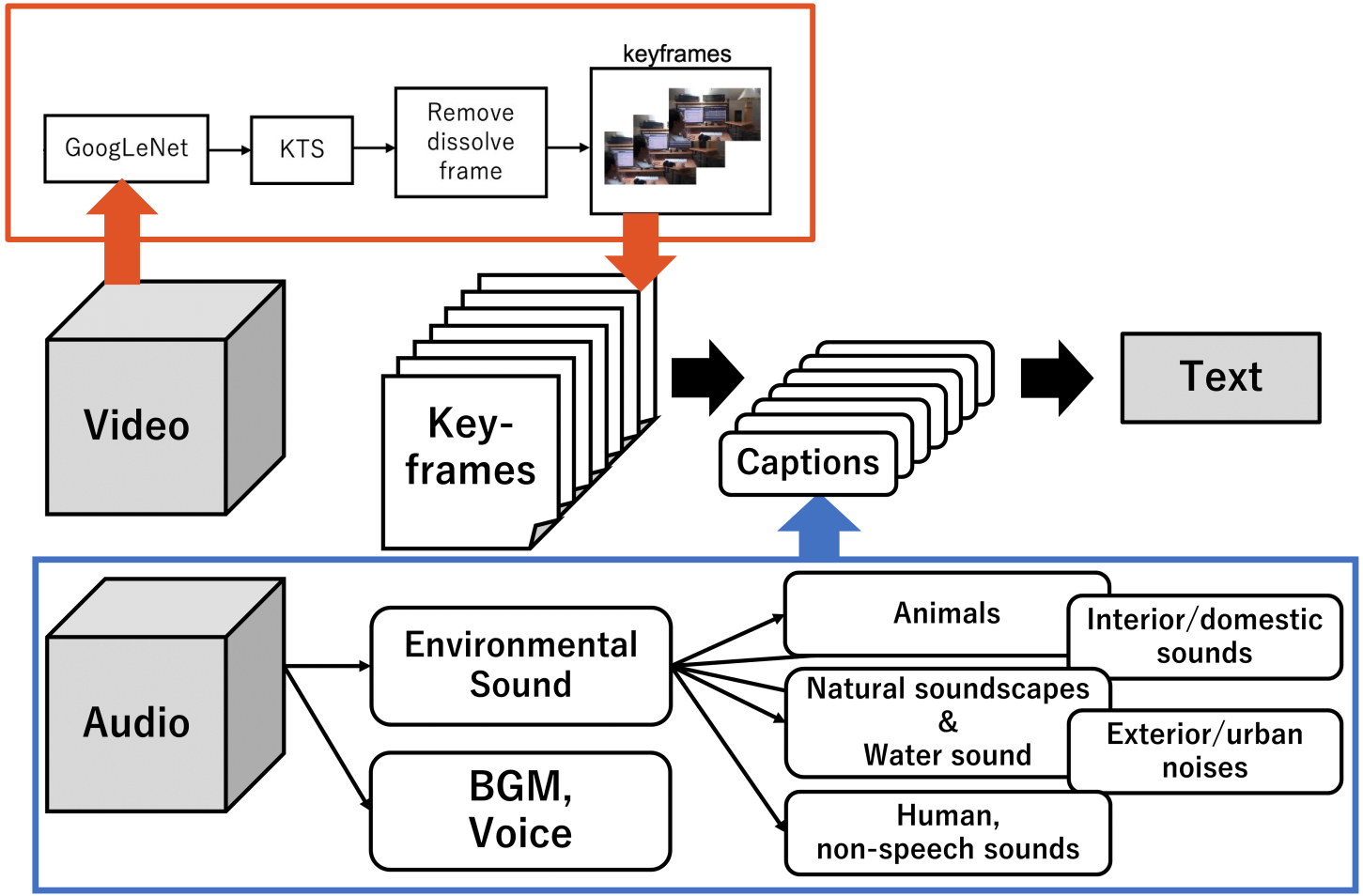
- Train data
  - ESC-50
- CNN
  - Mel spectrogram



Environmental Sound Classification Phase



- Labels and captions are vectorized using sentence-transformers, and calculate the Cosine similarity
- The similarity score is the Cosine similarity scaled to a range of 0 to 0.5
- Aligning sentence beginnings and proper nouns with uppercase letters


keyframes
GoogLeNet
KTS
Remove dissolve frame
Video
Key-frames
Captions
Text
Audio
Environmental Sound
BGM, Voice
Animals
Interior/domestic sounds
Natural soundscapes & Water sound
Exterior/urban noises
Human, non-speech sounds

Result

- tv23_NUT_1 and 3, which included the dissolve detection, scored higher than tv23_NUT_2 and 4 in the four metrics, excluding BLEU.

Run	Keyframe Extraction	Aggregation	METEOR	BLEU	CIDEr	CIDEr-D	spice
TV23_NUT_1	KTS + Dissolve Detection	Text	0.2274255377	0.0384961812	0.501	0.140	0.078
TV23_NUT_2	KTS	Text	0.2248083453	0.0392198399	0.484	0.130	0.076
TV23_NUT_3	KTS + Dissolve Detection	Text + Audio	0.2255115912	0.0539845496	0.495	0.139	0.077
TV23_NUT_4	KTS	Text + Audio	0.2232341071	0.0548268463	0.479	0.130	0.076

Observations: Keyframe Extraction

- The score increased compared to last year but improve is marginal.
- If only a portion of the video frame has been edited, the dissolve scene cannot be detected.

Run	Keyframe Extraction	Aggregation	METEOR	BLEU	CIDEr	CIDEr-D	spice
TV23_NUT_1	KTS + Dissolve Detection	Text	0.2274255377	0.0384961812	0.501	0.140	0.078
TV23_NUT_2	KTS	Text	0.2248083453	0.0392198399	0.484	0.130	0.076
TV23_NUT_3	KTS + Dissolve Detection	Text + Audio	0.2255115912	0.0539845496	0.495	0.139	0.077
TV23_NUT_4	KTS	Text + Audio	0.2232341071	0.0548268463	0.479	0.130	0.076

Observations: Environmental sound classification

- BLEU scores increasing due to the grammatical adjustments.
- The environmental sound classification is occasionally helpful in error handling but it is very rare.

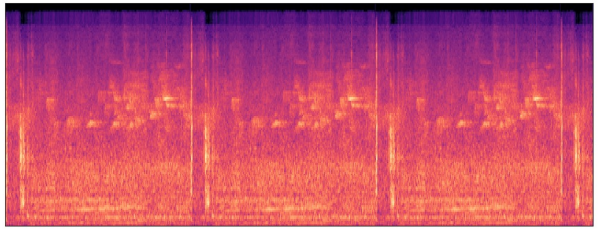
Run	Keyframe Extraction	Aggregation	METEOR	BLEU	CIDEr	CIDEr-D	spice
TV23_NUT_1	KTS + Dissolve Detection	Text	0.2274255377	0.0384961812	0.501	0.140	0.078
TV23_NUT_2	KTS	Text	0.2248083453	0.0392198399	0.484	0.130	0.076
TV23_NUT_3	KTS + Dissolve Detection	Text + Audio	0.2255115912	0.0539845496	0.495	0.139	0.077
TV23_NUT_4	KTS	Text + Audio	0.2232341071	0.0548268463	0.479	0.130	0.076

Observations: Environmental sound classification

【Video ID 484】

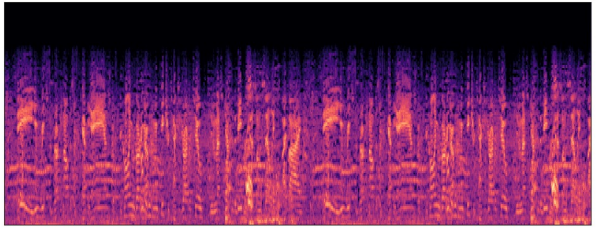
Natural soundscapes & water sound: Chirping birds



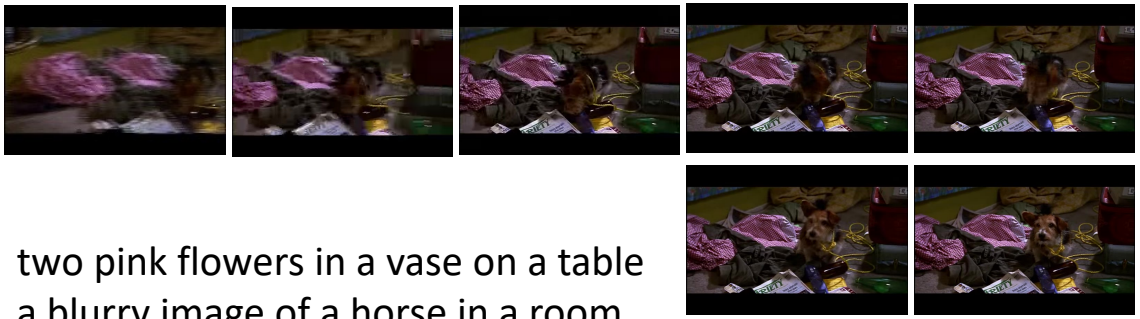



a young man laying on the ground in the grass
a man in a white shirt walking in the woods
a young man wearing a white tshirt standing in the woods
a man in a white shirt walking in the woods

Observations: Environmental sound classification



【Video ID 1156】 Animal: Cat



two pink flowers in a vase on a table
a blurry image of a horse in a room
a cat laying on a pile of clothes and money
a cat sitting on top of a pile of clothes
a pile of clothes and a cat on the floor
a dog sitting on top of a pile of clothes
a dog sitting in a pile of clothes on the floor

Conclusion

- **Keyframe Extraction Method:**
  - New keyframe extraction method is marginally more effective than previous approaches
  - Issues arise when the dissolve effect is applied only to a part of the screen, leading to a failure in removing dissolve frames.
- **Environmental Sound Classification Phase:**
  - The phase did not show improvement with the addition of the environmental sound classification phase.
  - Limiting the audio types and ensuring proper sound classification will enhance its effectiveness.