Multimodal 2DCNN action recognition from RGB-D Data with Video Summarization

Vicent Roig Ripoll

Master in Artificial Intelligence
UPC, UB, URV

Master's Thesis

Advisor: Sergio Escalera Guerrero
Co-advisor: Maryam Asadi-Aghbolaghi

October, 2017

Overview

Introduction

Motivation:

- Human action recognition research area
  - large intra-class variations
  - low video resolution
  - high dimension of video data
- Kinect $\rightarrow$ multimodal data access
- Hand-crafted features *vs* automatic feature learning

Goals:

- Analyse multimodal data benefits in deep learning
- To this end, 2DCNN is extended to multimodal (**MM2DCNN**)
- Evaluation of **video summarization** impact in action recognition

Outline

Hand-crafted Features

Approaches to cope with temporal information

1. Treat videos as spatio-temporal volumes
2. **Flow-based** features, explicitly deal with motion
3. **Trajectory-based** approaches, motion is implicitly modelled

- Histograms of Oriented Gradients (HOG) $\rightarrow$ HOG3D
- Scale-Invariant Feature Transform (SIFT) $\rightarrow$ 3D-SIFT
- Histogram of Normals (HON) $\rightarrow$ HON4D
- Dense Trajectories (DT & iDT)

Optical Flow

For a given time t and pixel $(x, y)_t$:

$$(x, y)_{t+1} = (x, y)_t + d_t^{(x,y)}$$

Applications:

- Trajectory construction
- Descriptors: HOF, MBH
- Deep learning $\rightarrow$ CNN input

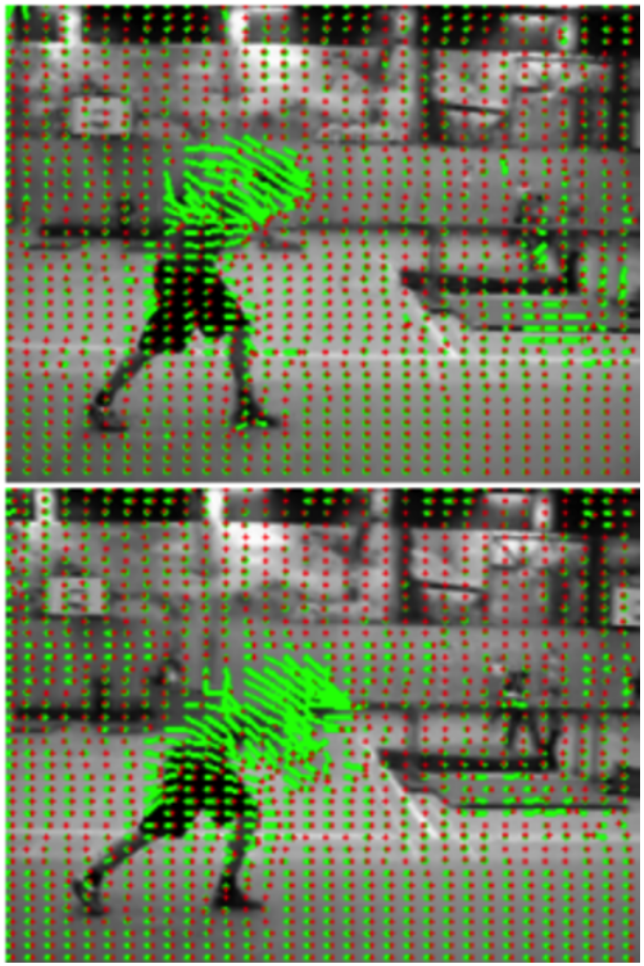


Figure: Optical flow field vectors (green vectors with red end points)

Scene Flow

For a given time t and pixel $(x, y, z)_t$

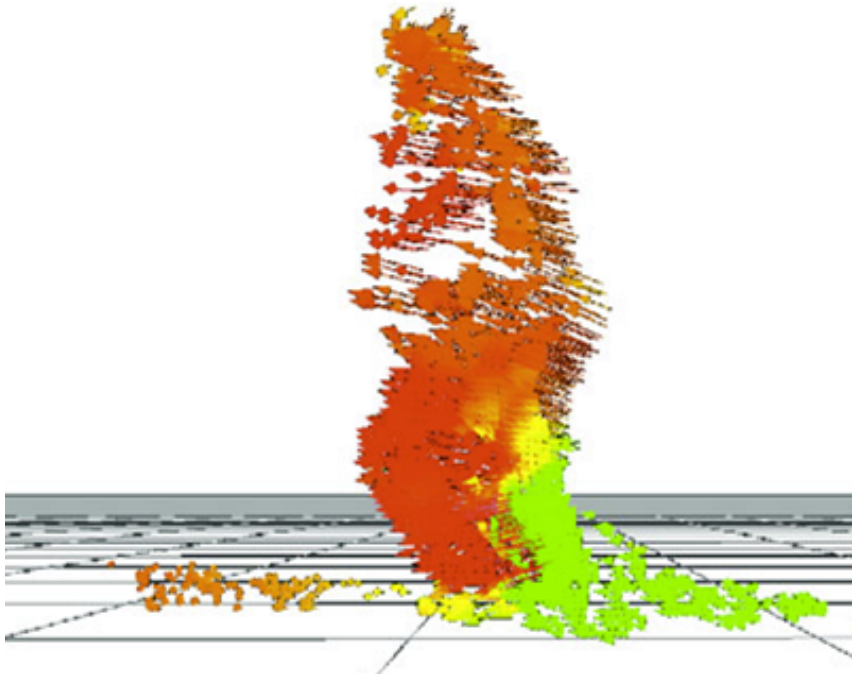
$$(x, y, z)_{t+1} = (x, y, z)_t + d_t^{(x,y,z)}$$

Applications:

- 3D trajectory construction
- Deep learning $\rightarrow$ CNN input

Advantages over optical flow:

- Real world motion units
- Z-axis motion



Deep Learning - Two-stream Convolutional Neural Network

2DCNN performs the recognition by processing 2 **different streams**, **spatial** and **temporal**, combining both by a late fusion

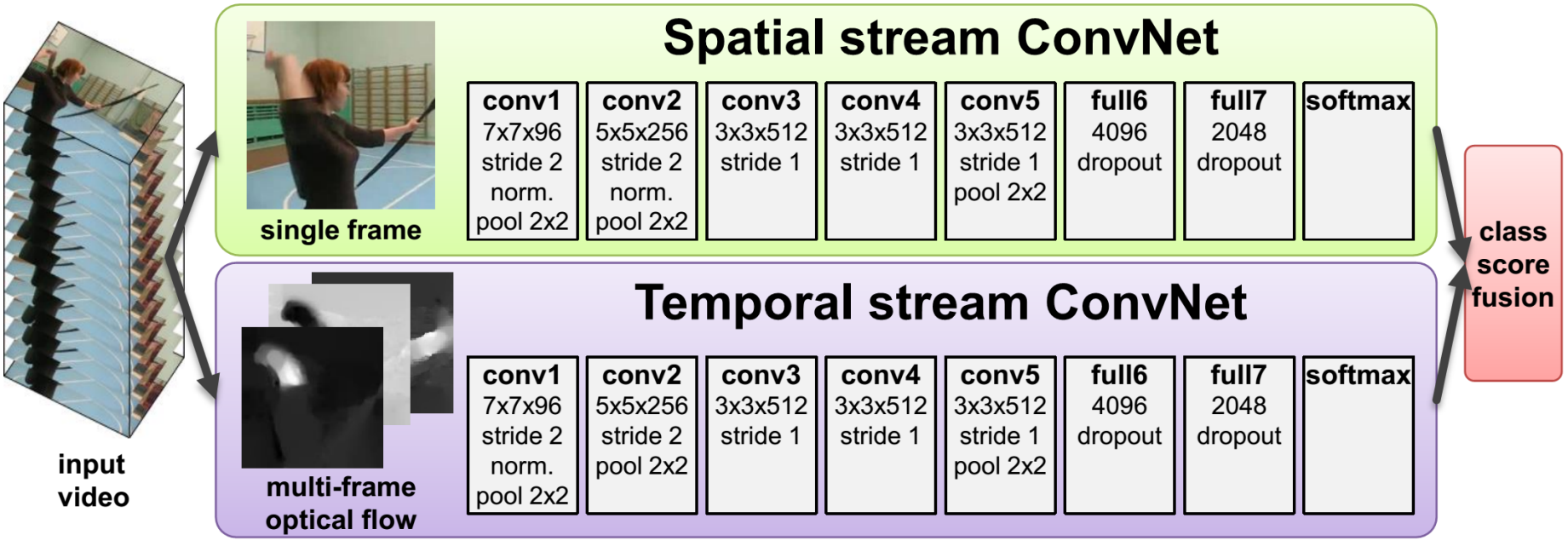


Figure: Two-stream architecture for video classification

Outline

Video Summarization

Video summarization allows for the extraction of few video frames (keyframes) so that they jointly try to **maximize** the **information** contained in the original video

Figure: Video summarization overview

Video Summarization - Techniques

Sequential Distortion Minimization (SeDiM) [Panagiotakis2013]

Selects frames so that the **distortion** between the original video and the synopsis is **minimized**. Does not guarantee global minima of distortion

Absolute Histogram Difference (Hdiff) [CV2015]

Simple summarization technique based on the **absolute difference** of histograms of consecutive frames

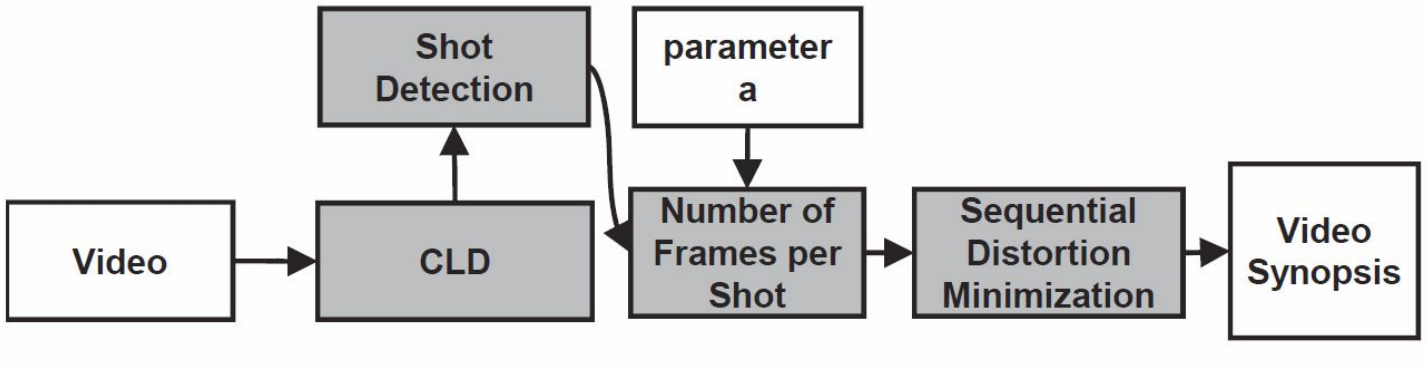
Time Equidistant Algorithm (TEA)

Keeps keyframes in **equal intervals** in duration

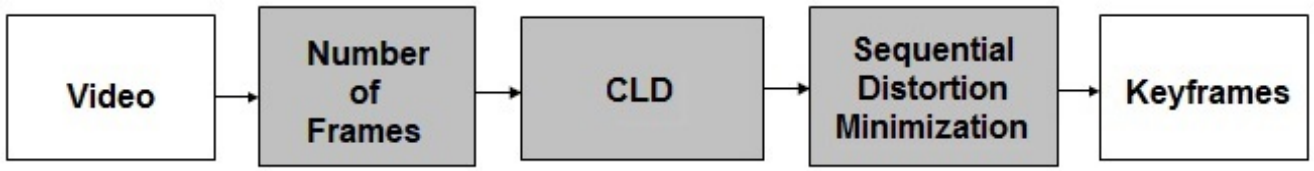
Content Equidistant Algorithm (CEA) [4783025]

Based on the *iso-content principle*. Estimates keyframes that are equidistant in **video content**

SeDiM - Architecture



(a) Original steps



(b) Modified version

Figure: Schemes for (a) **original** version and (b) our **proposal**

SeDiM - Examples

Figure: $k = 5$ keyframes on Montalbano RGB samples. 1st row: *vattene*, 2nd: *seipazzo*, 3th *combinato*, 4th: *ok*

Hdiff - Examples

Figure: $k = 5$ keyframes on Montalbano RGB samples. 1st row: *vattene*, 2nd: *seipazzo*, 3th *combinato*, 4th: *ok*

TEA - Examples

Figure: $k = 5$ keyframes on Montalbano RGB samples. 1st row: *vattene*, 2nd: *seipazzo*, 3th *combinato*, 4th: *ok*

CEA - Examples

Figure: $k = 5$ keyframes on Montalbano RGB samples. 1st row: *vattene*, 2nd: *seipazzo*, 3th *combinato*, 4th: *ok*

Outline

Proposed Method

1. Data Pre-processing

1. RGB-D Registering
2. Depth denoising

2. Video Summarization strategies

1. **RGB:** Ordered sequences of $k = 14$ **RGB videos**
2. **Depth:** Ordered sequences of $k = 14$ **Depth videos**
3. **RGB-D:** Combination of $k = 7$ **RGB** and **depth** summaries

3. Multi-Modal 2D CNN

- Extend **VGG-16** 2DCNN by adding a **scene flow stream**
- Base models are **UCF101** (temporal and spatial)
- Scene flow stream is to be fine-tuned from the RGB model of the same dataset
- **Weighted average** fusion

RGB-D Alignment

- Some datasets are not properly aligned
- **RGB-D registration** uses the **intrinsic** (focal length and the distortion model) and **extrinsic** (translation and rotation) camera parameters to warp the colour image to fit the depth map

Figure: IsoGD RGB and depth frame superpositions

Hybrid Median Filter

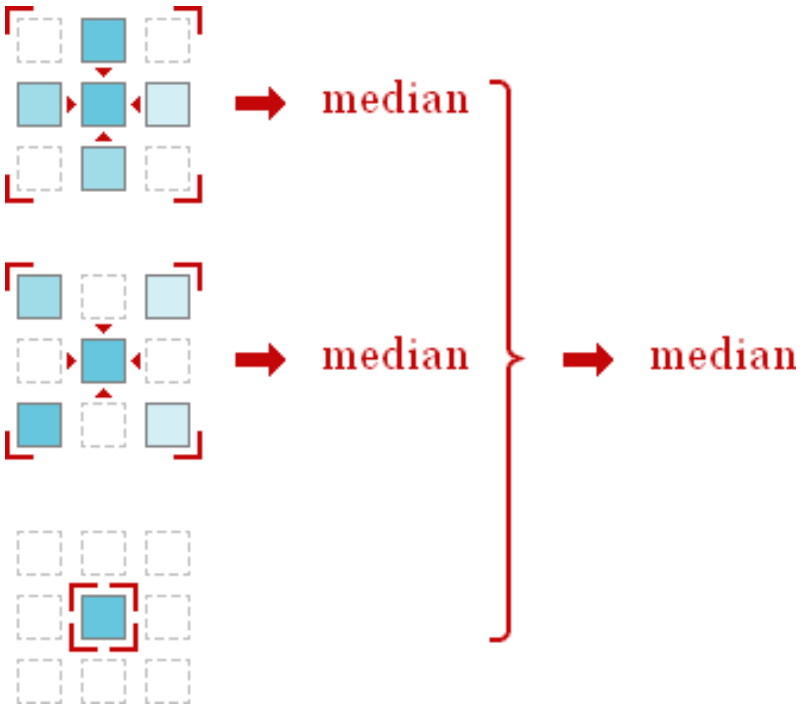


Figure: HMF workflow

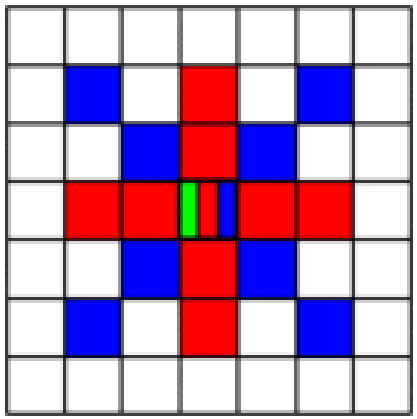


Figure: 5x5 HMF shapes

Denoising (1)

(a) Original

(b) Inpaint

(c) Inpaint + HMF

Denoising (2)

- 1st row: Inpainting + HMF
- 2nd row: Superposition **before** registration
- 3rd row: Superposition **after** registration

Late Fusion

Weighted sum is used to fuse class scores of each modality.

Given M modalities, each sample has N feature arrays of size K classes, then, the final scores are:

$$S_f = \sum_{i}^{N} w_i S_i$$

where weights w_i are to be optimized

Outline

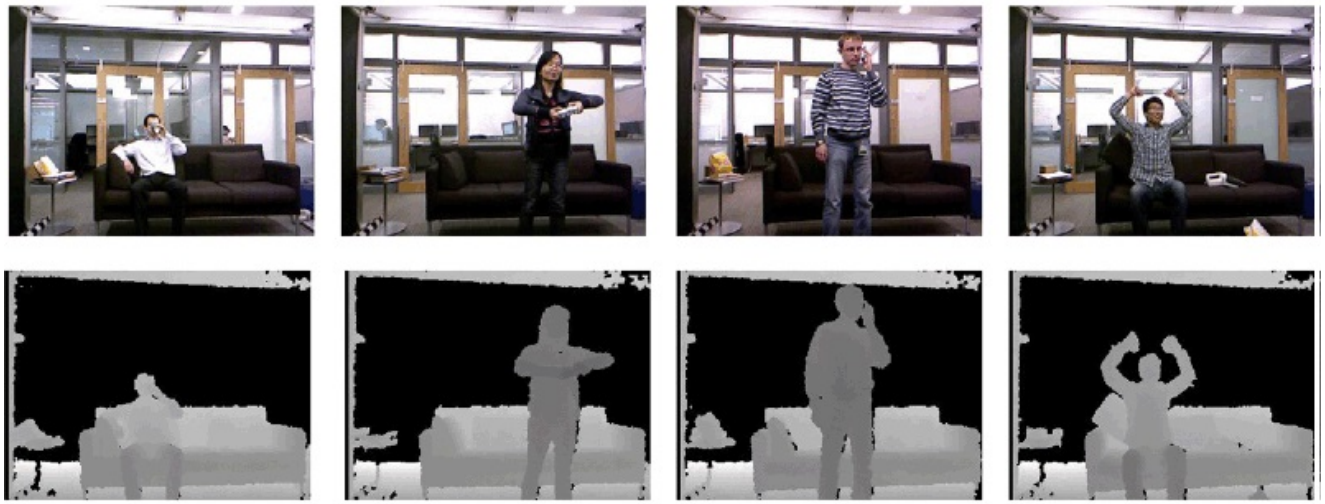
MSR Daily Activity 3D

Characteristics:

- Action recognition
- 16 classes
- 10 subjects
- 320 samples

Evaluation:

- 25% Train
- 25% Validation
- 50% Test



Montalbano V2

Characteristics:

- Gesture recognition
- 20 classes
- 27 subjects
- 940 samples
- 13858 gestures

Evaluation:

- 1-470 Train
- 471-700 Validation
- 701-940 Test

Isolated Gesture Dataset (IsoGD)

Characteristics:

- Gesture recognition
- 249 classes
- 17 subjects
- 47933 gestures

Evaluation:

- 35878 Train
- 5784 Validation
- 6271 Test

Evaluation on MSR Daily

Model	RGB	Depth	RGB-D	Random
RGB	53.75	54.37	53.75	52.50
Opt. flow	55.63	52.50	55.00	53.75
Scene flow	62.50	59.84	61.42	62.20
Late Fusion	**67.72**	64.57	63.78	66.14

Figure: sedim

Model	RGB	Depth	RGB-D	Random
RGB	51.25	51.88	52.50	52.50
Opt. flow	53.12	53.12	55.00	53.75
Scene flow	59.84	59.84	56.69	62.20
Late Fusion	67.72	**68.50**	**68.50**	66.14

Figure: tea

Model	RGB	Depth	RGB-D	Random
RGB	51.88	50.62	51.88	52.50
Opt. flow	51.25	48.75	51.88	53.75
Scene flow	58.27	55.12	58.27	62.20
Late Fusion	62.20	63.78	65.35	**66.14**

Figure: hdiff

Model	RGB	Depth	RGB-D	Random
RGB	53.12	53.75	54.37	52.50
Opt. flow	56.87	53.12	54.37	53.75
Scene flow	59.84	57.48	59.06	62.20
Late Fusion	66.14	66.14	**66.93**	66.14

Figure: cea

W=[0.2, 0.3, 0.5]

Evaluation on Montalbano V2

Model	RGB	Depth	RGB-D	Random
RGB	96.03	97.06	95.72	97.06
Opt. flow	61.06	59.74	60.67	64.24
Scene flow	66.79	66.31	65.15	64.72
Late Fusion	97.28	**97.56**	97.20	97.25

Figure: sedim

Model	RGB	Depth	RGB-D	Random
RGB	97.51	97.48	97.37	97.06
Opt. flow	63.09	62.69	63.63	64.24
Scene flow	70.41	70.33	68.60	64.72
Late Fusion	97.68	97.70	**97.74**	97.25

Figure: tea

Model	RGB	Depth	RGB-D	Random
RGB	96.10	96.47	96.10	97.06
Opt. flow	60.24	62.31	59.79	64.24
Scene flow	66.03	67.21	65.63	64.72
Late Fusion	96.46	96.89	96.27	**97.25**

Figure: hdiff

Model	RGB	Depth	RGB-D	Random
RGB	96.50	96.77	96.55	97.06
Opt. flow	62.15	63.34	62.73	64.24
Scene flow	66.82	67.13	66.96	64.72
Late Fusion	96.94	97.19	97.17	**97.25**

Figure: cea

W=[0.65, 0.15, 0.2]

Evaluation on IsoGD

Model	RGB	Depth	RGB-D	Random
RGB	26.37	27.32	27.47	21.50
Opt. flow	39.82	39.58	39.88	42.74
Late Fusion	42.95	43.43	**43.69**	43.66

Figure: sedim

Model	RGB	Depth	RGB-D	Random
RGB	30.93	30.38	30.24	21.50
Opt. flow	41.67	41.34	41.60	42.74
Late Fusion	**46.62**	**46.63**	46.35	43.66

Figure: tea

Model	RGB	Depth	RGB-D	Random
RGB	27.11	29.36	27.48	21.50
Opt. flow	40.11	39.79	39.35	42.74
Late Fusion	42.93	**44.56**	43.09	43.66

Figure: hdiff

Model	RGB	Depth	RGB-D	Random
RGB	26.68	29.29	27.98	21.50
Opt. flow	41.10	42.28	41.95	42.74
Late Fusion	43.52	**45.91**	45.08	43.66

Figure: cea

W=[0.2, 0.8]

Comparison - MSR Daily

Method	Accuracy
EigenJoints	58.10
MovingPose	73.80
HON4D	80.00
SSTKDes	85.00
ActionLet	**85.75**
MMDT	78.13
MM2DCNN	68.50

Table: Performance comparison with *sota* methods on MSR Daily

Comparison - Montalbano V2

Method	Accuracy
Rank pooling	75.30
AdaBoost, HoG	83.40
Temp Conv + LSTM	94.49
Dense Trajectories	83.50
MMDT	85.66
MM2DCNN	**97.74**

Table: Performance comparison with *sota* methods on Montalbano

Comparison - IsoGD

Method	Accuracy
NTUST	20.33
MFSK	24.19
MFSK+DeepID	23.67
XJTUfx	43.92
XDETVP-TRIMPS	50.93
TARDIS	40.15
ICT NHCI	46.80
AMRL	55.57
2SCVN-3DDSN	**67.19**
MM2DCNN	46.63

Table: Performance comparison with *sota* methods on IsoGD

ref: http://chalearnlap.cvc.uab.es

Multimodal Fusion Justification

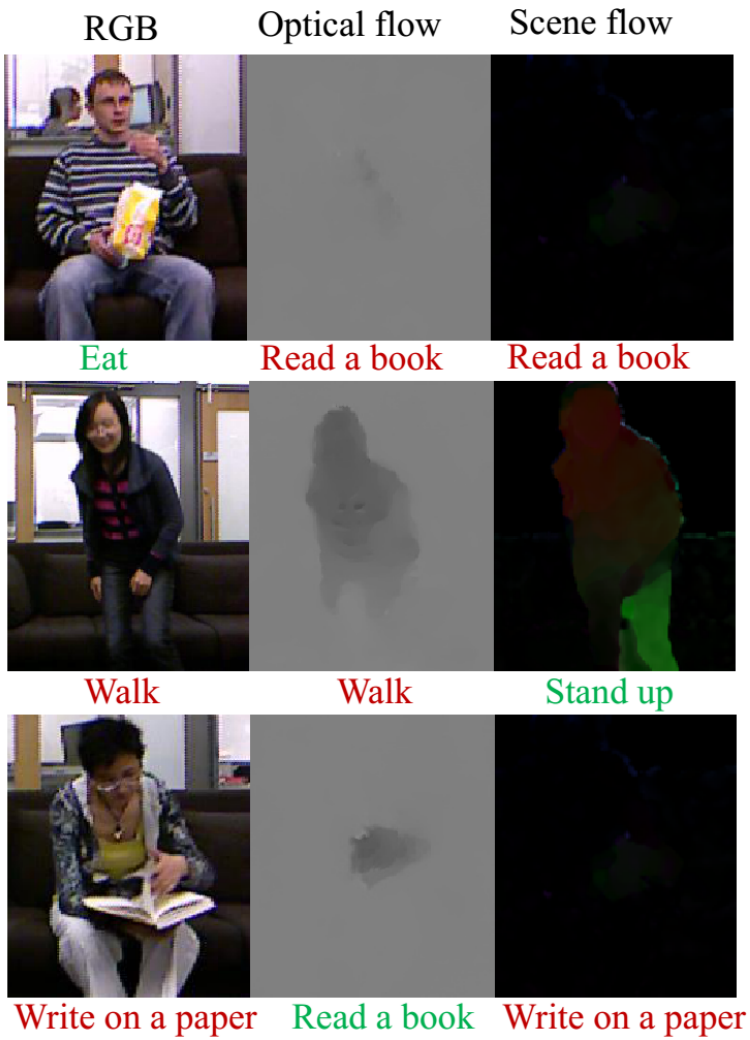


Figure: Each column shows one modality. Each row shows the result of each modality. Red: wrong, Green: correct

Outline

Conclusions

- Different summarization strategies do not change much
- TEA gets best results in all datasets
- State of the art in Montalbano V2
- MM2DCNN outperforms 2DCNN

Future Work

Video Summarization

- Different k per dataset
- Consider **other** video summarization alternatives

MM2DCNN

- Add scene flow stream for IsoGD
- Use a **larger dataset** (e.g. NTU) to pre-train all nets
- Fuse by using a trained model (Multiclass-SVM, RF, etc)
- **Apply PCA** to avoid overfitting

Others

- Combine **hand-crafted** features with **deep learning** [ICC2]
- Use **3DCNN** instead of 2DCNN

References

Panagiotakis, Costas and Ovsepian, Nelly and Michael, Elena

Video Synopsis Based on a Sequential Distortion Minimization Method

Computer Analysis of Images and Patterns: 15th International Conference

C. Panagiotakis and A. Doulamis and G. Tziritas

Equivalent Key Frames Selection Based on Iso-Content Principles

IEEE Transactions on Circuits and Systems for Video Technology

Asadi-Aghbolaghi, Maryam and Bertiche, Hugo and Roig, Vicent and Kasaei, Shohreh and Escalera, Sergio (2017)

Action Recognition From RGB-D Data: Comparison and Fusion of Spatio-Temporal Handcrafted Features and Deep Strategies

The IEEE International Conference on Computer Vision (ICCV)

C V, Sheena and Narayanan, N.K.

Key-frame Extraction by Analysis of Histograms of Video Frames Using Statistical Methods

Procedia Computer Science (2015)

Thanks for your attention

Questions?