Maryam Asadi-Aghbolaghi[1], Hugo Bertiche[2], Vicent Roig[2], Shohreh Kasaei[1], Sergio Escalera[2,3]
[1]Dept. of Computer Engineering, Sharif University of Technology, Tehran, Iran
[2]Dept. of Applied Mathematics and Analysis, University of Barcelona, Barcelona, Spain
[3]Comuter Vision Center, Autonomous University of Barcelona, Spain
Email: masadia@ce.sharif.ir

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

Abstract

In this work,

- **Multimodal** fusion of RGB-D data are analyzed for action recognition by using **scene flow** as **early fusion** and integrating the results of all modalities in a **late fusion** fashion.
- **Multimodal dense trajectory (MMDT)** is proposed to describe RGB-D videos as handcrafted features.
- **Multimodal 2D CNN (MM2DCNN)** is proposed as the extension of 2D CNN by adding one more input stream (scene flow).
- The proposed methods are evaluated on **two action datasets**.
- Fusion of handcrafted and learning-based features achieved the state of the art results.

Introduction

- ❑ **Action recognition** is an active research area with potential applications of health-care monitoring, interactive gaming, surveillance, and robotics.
- ❑ **Microsoft Kinect** have facilitated capturing of low-cost depth images in real-time alongside color images (**multimodal** data).
  - **Late fusion** of RGB, depth, and motion-based representations (like optical flow) is an effective method for action recognition.
- ❑ **Scene flow [1]** is the real 3D motion of objects that move completely or partially with respect to a camera.
  - ✓ Considered as **Early fusion** of RGB and depth,
    - Preserving 3D motion data on the spatial structure of both modalities,
  - ✓ More discriminative than optical flow,
    - When it is significant **motion perpendicular to the image plane**,
  - ✓ **Invariant to the distance between objects and the camera**.
    - In 3D world, distance between two objects does not depend on the relative position to the camera while the same movement performed at different position may produce different optical flow in terms of pixels.

Multimodal Data

- ❑ **MMDT** is presented as a **handcrafted representation**.
  - ➢ *Dense trajectories* (DT) [2], **pruned** by exploiting scene flow data,
  - ➢ *Histogram of normal vector* (**HON**) is extracted from normal vectors of depth images.
- ❑ **MM2DCNN** is presented as **learning-based features**.
  - ➢ By the **incorporation of scene flow** information as a new model.
  - ➢ **Late fusion**: **score averaging** of the result of multi streams 2DCNN [3,4] (RGB, optical flow, and scene flow)
- ❑ Second fusion: **combination of handcrafted and deep models**,
  - ✓ Handcrafted: powerful in describing **motion information**,
  - ✓ Deep learning: good at describing **appearance data**.

Denoising and RGB-D Alignment

❑ **Denoising**

Missing pixels in depth images due to:

- × Limitations of the **IR sensor**,
- × Special **reflectance materials**,
- × **Distance** from the objects to the camera.

- ✓ **Interpolating** zero value pixels by its surrounding data,
- ✓ ***Hybrid median filter*** **(HMF)** to reduce pixel flickering,
  - Compute medians for different spatial directions
    - **Horizontal/vertical + diagonal**
  - Compute the **median of both** of them

❑ **RGB-D alignment**

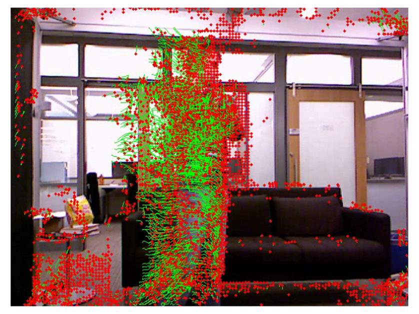
- × IR and optical cameras are **separated**,
- ✓ **Warp the color image to fit the depth one**,
  - Use the **intrinsic** (focal length and the distortion model) and **extrinsic** (translation and rotation) camera parameters.

Denoising and RGB-D Alignment

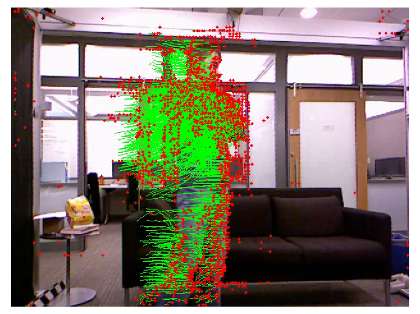
Multimodal Dense Trajectory (MMDT)

Trajectories

- ❑ Compute **scene flow** along the trajectories,
- ❑ **Pruning** dense trajectories,
  - By the information achieved by scene flow in meters.
- ✓ Scene flow is **invariant to the position of the subject relative to the camera**,
- ✓ Scene flow has an **additional dimension**, which allows the measurement of motion through **Z-axis**.



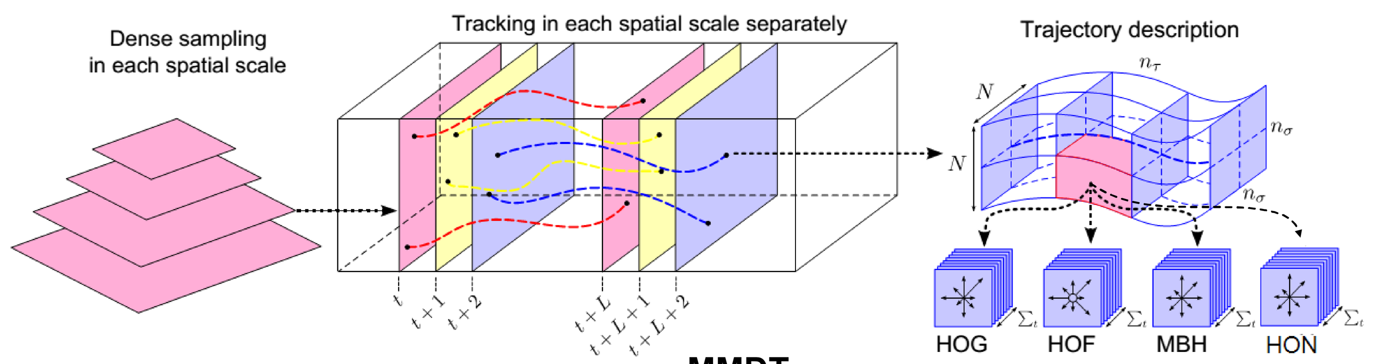
Without pruning



With pruning

HON descriptor

- ❑ New source of information; i.e., **depth maps**.
- ❑ Each normal is represented by **two angles** θ and φ:
  - $0 < \theta < \pi$ and $-\pi/2 < \varphi < \pi/2$,
- ❑ 5 bins are considered, (size of $\pi/4$ radians), total of 25 bins for sub-histogram,
- ❑ The final descriptor is the **concatenation** of 12 sub-histograms results in 300 dimensions.



MMDT

Video Summarization

- × Deep methods mostly select **a fixed number of frames with equal temporal spacing** between them. Thus, some relevant information might be lost.
- ✓ **Key frames selection**
  - Select relevant visual information to discriminate actions,
  - Keeping the size of the data small.

❑ ***Sequential Distortion Minimization*** **(SeDiM) [4]**

- The distortion between the original video and the synopsis video is minimized,
- Computationally feasible and discriminative way to extract key frames.

Key frames of three samples

Multimodal 2D CNN (MM2DCNN)

Three streams with 2D CNN (VGG-16)

- ❑ **Spatial network (RGB)**
  - Operating on key frames,
  - Using a pre-trained network on UCF-101.
- ❑ **Temporal network (Optical flow)**
  - Using volumes of stacking optical flow fields between several consecutive frames,
  - Using a pre-trained network on UCF-101.
- ❑ **Temporal network (Scene flow)**
  - Consider three dimensions of scene flow as three input channels,
  - Using a pre-trained model of its own RGB model.

Experimental Result

MSR Daily Dataset:

MMDT:

Table 1: DT and MMDT accuracy on MSRDaily Act. 3D.

Descriptors	DT	MMDT
HOG (RGB)	43.125	45.625
HON (Depth)	-	72.5
HOF + MBH (Opt. flow)	62.5	70
Best	**63.125**	**78.13**

Montalbano II:

Table 2: DT and MMDT accuracy on Montalbano II.

Descriptors	DT	MMDT
HOG (RGB)	67.3	67.3
HON (Depth)	-	77.67
HOF + MBH (Opt. flow)	82.0	82.0
Best	**83.5**	**85.66**

MM2DCNN:

Table 3: Accuracy for SeDiM on MSR Daily Activity 3D.

Model	RGB	Depth	RGB-D	Random
RGB	53.91	53.12	53.91	53.12
Opt. flow	55.47	57.81	55.47	55.70
Scene flow	67.19	68.75	66.41	64.84
Late Fusion	70.08	**71.65**	70.08	69.29

Table 4: Accuracy for SeDiM on Montalbano II.

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	69.90	69.68	69.02	70.93
Late Fusion	96.28	96.25	96.16	**97.06**

Second Late Fusion of MMDT and MM2DCNN:

Table 5: Second late fusion of MMDT and MM2DCNN.

Dataset	Accuracy
MSR Daily	82.50
Montalbano II	97.44

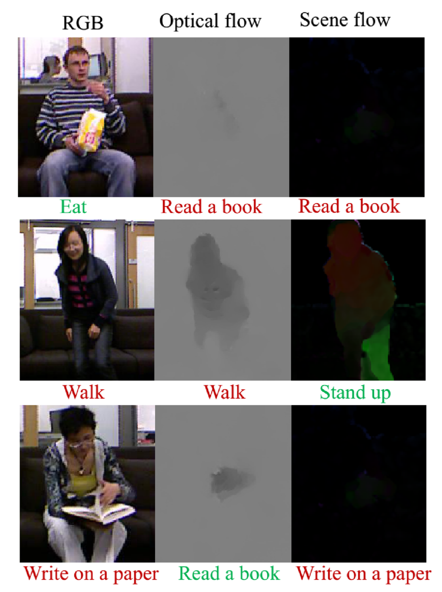
Comparison:

Table 6: Performance comparison on MSRDaily Act. 3D.

Method	Accuracy
EigenJoints[43]	58.10
MovingPose[44]	73.80
HON4D [15]	80.00
SSTKDes [16]	85.00
ActionLet [40]	**85.75**
MMDT	**82.50**
MM2DCNN	71.65

Table 7: Performance comparison on Montalbano II.

Method	Accuracy/Precision
Fernando et al. [45]	75.3
Pigou et al. [46]	94.49
MMDT	85.66
MM2DCNN	**97.44 (97.52 Precision)**



Examples from MSR Daily. Each column shows one modality. Each rows shows the classification result of each modality.
Red: Wrong classification, Green: Correct classification.

References

[1] Mariano Jaimez, Mohamed Souiai, Javier GonzalezJimenez, and Daniel Cremers. A primal-dual framework for real-time dense rgb-d scene flow. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 98–104. IEEE, 2015.

[2] Heng Wang, Alexander Klaser, Cordelia Schmid, and ¨ Cheng-Lin Liu. Action recognition by dense rajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

[3] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576. 2014.

[4] Limin Wang, Xiong Yuanjun, Wang Zhe, and Qiao Yu. "Towards good practices for very deep two-stream convnets." *arXiv preprint arXiv:1507.02159* (2015).]

[5] Costas Panagiotakis, Nelly Ovsepian, and Elena Michael. Video synopsis based on a sequential distortion minimization method. In *International Conference on Computer Analysis of Images and Patterns*, pages 94–101. Springer, 2013.