SPATIO-TEMPORAL FASTMAP-BASED MAPPING FOR HUMAN ACTION RECOGNITION

Lilia Chorfi Belhadj and Max Mignotte
Département d'Informatique et de Recherche Opérationnelle (DIRO), Université de Montréal
Montréal, Canada

ABSTRACT

This paper presents a simple and efficient method for action recognition based on the learning of an explicit representation for an intrinsic dynamic shape manifold of human action. The proposed model relies on a short temporal set of FastMap dimensionality reduction-based technique for embedding a sequence of raw moving silhouettes, associated to an action video into a low-dimensional space, in order to characterize the spatio-temporal property of the action, as well as to preserve much of the geometric structure. The objective is to provide a recognition method that is both simple, fast and applicable in many scenarios. Moreover, we demonstrate the robustness of our method to partial occlusion, deformation of shapes, significant changes in scale and viewpoint, irregularities in the performance of an action, and low-quality video.

Introduction

- **Human activity recognition:**
  - **Local representations**
  - **Global representations:**
    - Spatio-temporal template matching
    - Manifold learning based
- **Our solution:** hybrid-framework which combines both manifold learning and spatio-temporal template matching technique

Spatio-Temporal Action Volume (STV)

Build a centered motion field of a moving body by aligning the 2D center of mass corresponding to each normalized binary silhouettes to a reference point:

- To represent how as opposed to where
- Global translational speed of the movement is less informative than the motion of body parts relative to the torso of the person over time

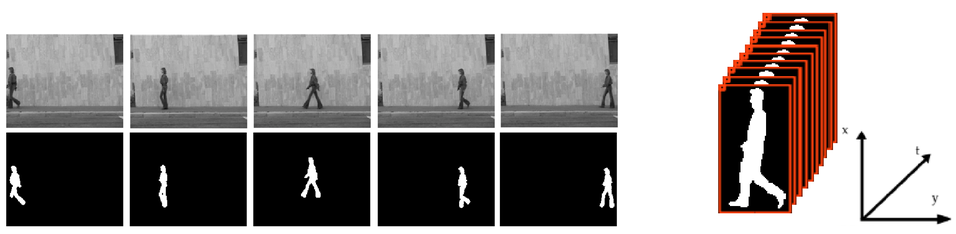
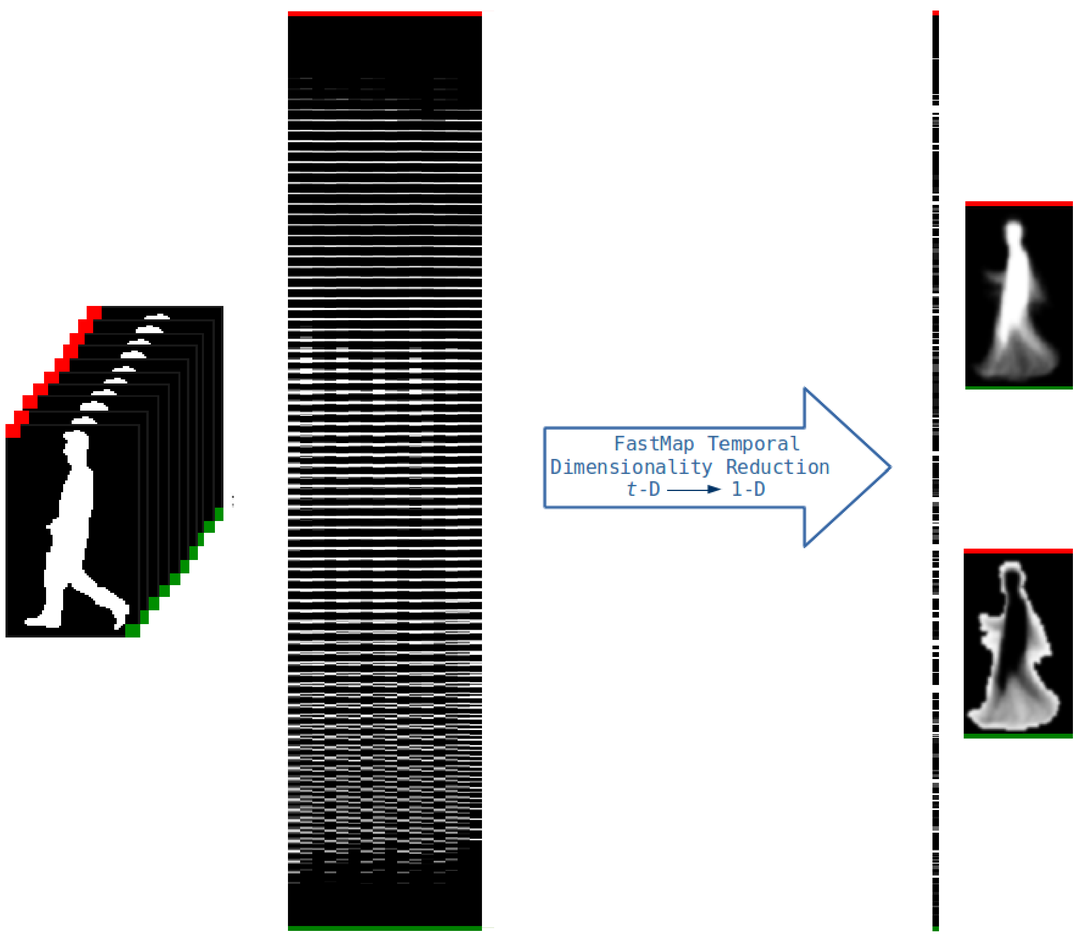


Figure : Generating Spatio-Temporal Action Volume (STV)

Spatio-Temporal Action Shapes (STAS)

- To treat both periodic and non-periodic actions and to compensate for different lengths of the sequences:
  - Use a temporal sliding window of 10 frames with an ovelap of 5 frames along the temporal axis.
- Divide STV into $H \times W$ vectors
  - (H, W) are, respectively, height and width of the image
  - Each pixel-vector has a dimension T (number of frames in the STV)
  - Each pixel-vector contains the intensity values that the pixel takes over T consecutive frames of a video sequence.
- FastMap mapping from TD to $1D$ along the temporal axis
  - Each pixel-vector in the initial space corresponds to a point in the reduced space
  - These points will represent the pixel intensity values of the STAS image



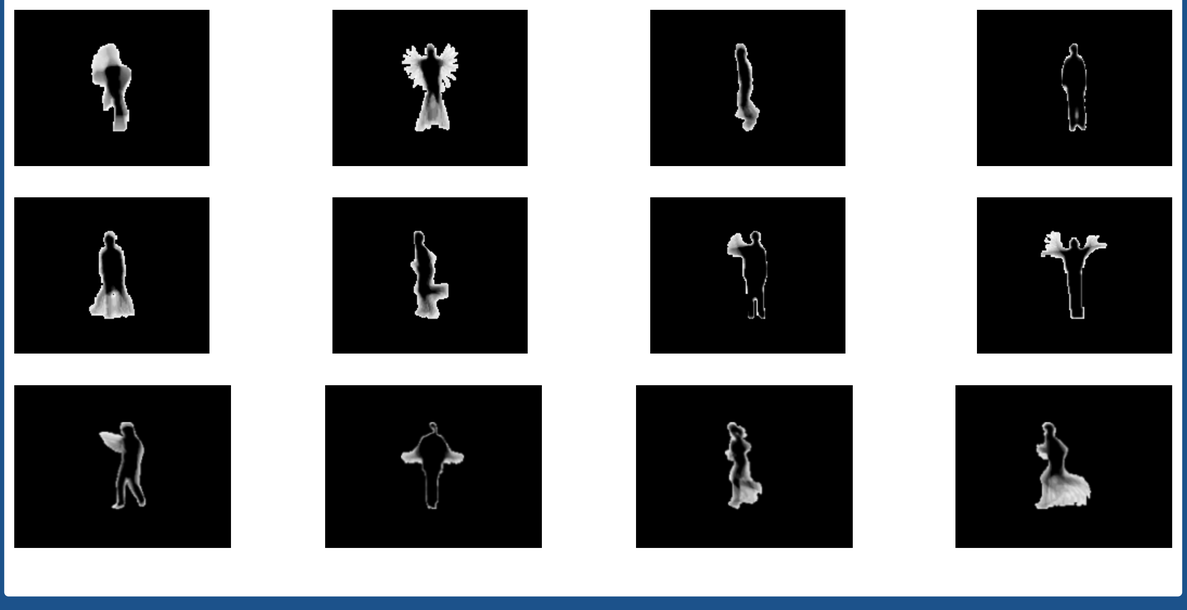
- Mean correlation rates for the FastMap reduction technique:
  - estimate the correlation ρ of the Euclidean distance between each pairwise vectors-pixels in the high dimensional space (let X be this vector) and their corresponding Euclidean distances in 1-dimensional space (let Y be this vector) following equation:

$$\rho_{X,Y} = corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y} = \frac{X^t Y / |X| - \bar{X}\bar{Y}}{\sigma_X \sigma_Y} \quad (1)$$

  where X^t, $|X|$, $\bar{X}$ and σ_X respectively represent the transpose, cardinality, meann and standard deviation of X.

Actions	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walkk	Wave1	Wave2
Correlation	0.79	0.88	0.86	0.92	0.74	0.89	0.80	0.90	0.50	0.82

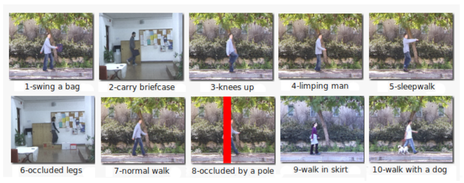
Sample images representing the different action STAS



Experimental Results

- Use a Nearest Neighbors Classifier with leave-one-out procedure to label the test actions
- Recognition results on the Weizmann dataset:

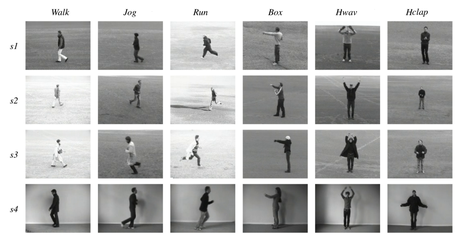
method [1]'14	NNC	92.3%
method [2]'08	NNC	95.56%
method [3]'15	NNC	96.3%
method [4]'07	NNC	97.83%
method [5]'15	SVM	99.1%
method [6]'15	SVM	100%
method [7]'15	SVM	100%
Our method	**FastMap**	**100%**



Actions	1	2	3	4	5	6	7	8	9	10
Corrupted dataset										
method [4]	8	8	8	8	8	8	8	8	8	8
method [2]	**2**	8	**6**	**2**	8	8	8	**6**	8	**7**
method [8]	8	8	8	8	**3**	8	8	8	8	**5**
method [9]	8	8	8	8	**2**	8	8	8	8	**5**
Our method	8	8	8	8	8	8	8	8	8	8

Actions	1	2	3	4	5	6	7	8	9	10
Changing of viewpoints dataset										
method [4]	8	8	8	8	8	8	8	8	8	8
method [2]	8	8	8	8	8	8	**6**	**10**	**2**	**2**
method [9]	8	8	8	8	8	8	8	**2**	**5**	**2**
Our method	8	8	8	8	8	8	8	8	**6**	**6**

- Recognition results on the KTH dataset:



	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	100	0	0	0	0	0
Handclapping	0.75	99.24	0	0	0	0
Handwaving	0	0.26	99.74	0	0	0
Jogging	0	0.25	0	83	5.75	11
Running	0	3.01	0	26.44	71.53	3.01
Walking	0.5	0	0	0.75	0	98.75

	Representation	Accuracy
method [10]'10	low-level	82%
method [11]'08	low-level	84.3%
method [12]'10	low-level	87.3%
method [13]'10	low-level	90.57%
method [14]'15	low-level	92.13%
method [6]'15	low-level	93.98%
method [5]'15	low-level	95.8%
method [15]'10	high-level	94.5%
method [16]'12	high-level	98.9%
method [17]'14	high-level	99.54%
Our method	**low-level**	**92.04%**

Conclusion

Our method:

- is simple, inexpensive and fast allowing simple action recognition.
- does not require video alignment, and is applicable in many scenarios.
- is robust to partial occlusion, deformation of shapes, significant changes in scale and viewpoint, irregularities in the performance of an action, and low-quality video.

References

[1] R. Touati and M. Mignotte, "Mds-based multi-axial dimensionality reduction model for human action recognition," in *Computer and Robot Vision (CRV)*, pp. 262–267, IEEE, 2014.
[2] H. Ragheb, S. Velastin, P. Remagnino, and T. Ellis, "Human action recognition using robust power spectrum features," in *Image Processing ICIP*, pp. 753–756, IEEE, 2008.
[3] T. Zhang, L. Xu, J. Yang, P. Shi, and W. Jia, "Sparse coding-based spatiotemporal saliency for action recognition," in *Image Processing ICIP*, pp. 2045–2049, IEEE, 2015.
[4] L. Gorelick, M. Galun, E. Sharon, R. Basri, and A. Brandt, "Shape representation and classification using the poisson equation," *Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1991–2005, 2006.
[5] K. Xu, X. Jiang, and T. Sun, "Human activity recognition based on pose points selection," in *Image Processing ICIP*, pp. 2930–2834, IEEE, 2015.
[6] Y. Shao, Y. Guo, and C. Gao, "Human action recognition using motion energy template," *Optical Engineering*, vol. 54, no. 6, pp. 063107–063107, 2015.
[7] D. Vishwakarma, R. Kapoor, and A. Dhiman, "A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics," *Robotics and Autonomous Systems*, 2015.
[8] L. Wang and D. Suter, "Visual learning and recognition of sequential data manifolds with applications to human movement analysis," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 153–172, 2008.
[9] F. Zheng, L. Shao, and Z. Song, "A set of co-occurrence matrices on the intrinsic manifold of human silhouettes for action recognition," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 454–461, ACM, 2010.
[10] J. Yin and Y. Meng, "Human activity recognition in video using a hierarchical probabilistic latent model," in *Computer Vision and Pattern Recognition Workshops*, pp. 15–20, IEEE, 2010.
[11] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference*, pp. 275–1, British Machine Vision Association, 2008.
[12] W. Yang, Y. Wang, and G. Mori, "Recognizing human actions from still images with latent poses," in *Computer Vision and Pattern Recognition*, pp. 2030–2037, IEEE, 2010.
[13] M. Kaâniche and F. Bremond, "Gesture recognition by learning local motion signatures," in *Computer Vision and Pattern Recognition*, pp. 2745–2752, IEEE, 2010.
[14] A. Iosifidis, A. Tefas, and I. Pitas, "Merging linear discriminant analysis with bag of words model for human action recognition," in *Image Processing ICIP*, pp. 832–836, IEEE, 2015.
[15] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Computer Vision and Pattern Recognition*, pp. 2046–2053, IEEE, 2010.
[16] S. Sadanand and J. Corso, "Action bank: A high-level representation of activity in video," in *Computer Vision and Pattern Recognition*, pp. 1234–1241, IEEE, 2012.
[17] A. Iosifidis, A. Tefas, and I. Pitas, "Regularized extreme learning machine for multi-view semi-supervised action recognition," *Neurocomputing*, vol. 145, pp. 250–262, 2014.

chorfibl@iro.umontreal.ca
mignotte@iro.umontreal.ca