Sharif University of Technology
Department of Computer Engineering

Master of Science Thesis

Human Action Recognition Using Spatiotemporal Features

By:
Amir Ghodrati

Supervisor:
Dr. Shohreh Kasaei

January 2010

Introduction

In this thesis, a broad study on human action recognition is done and some techniques to improve state of the art results are developed. The thesis is covered by these chapters: related works, proposed methods, evaluation and experimental results, conclusion and future works.

Related Works

In this chapter, different methods and techniques which I have studied during my master period are classified. During this chapter, the task of action recognition is divided to Motion representation and classification. As I found, representation is more important part than classification. Gaining more discrimination power during representation leads to using less expensive classifiers (such as SVM, K-NN).

Motion representations can be categorized to parametric representations, global representations and local representations. Each of which have some pros and cons that illustrated in table 1.

Table 1: comparison of different representations.

Models	**pros**	**cons**	**tips**
Parametric representation	Psychological approach, industrial applications in medics and making animations.	Finding parts of body, parameter estimation for optimization, depending to tracking, heavy interaction with user	These approaches are used just in controlled settings and are not applied to realistic actions
Global representation	Invariant to color and texture, more easy representation related to parametric models, suitable for recognition actions at a distance.	Depend to background subtraction or optical flow computations, sensitive to view point	
Local representation	Hybrid of parametric and global representations, good results, robust to clutter, not dependent to background subtraction	Do not model geometrics of action, heavy feature matching	Applied to uncontrolled setting[1] but does not handle camera motions

Due to significant advantages of local representations, this approach is used in this thesis but however a broad comparison (implementation) in same bedrock, using realistic actions in movies and sport, should be done.

We study four types of local space-time feature detectors. 3D Harris developed by Laptev[2] which supports automatic scale selection, Cuboids developed by Dollar[3] which produce rich set of features, Volumetric features developed by Ke[4] which have efficient computations, Salient space-time features developed by Oikonomopoulos[5] which is inspired by Kadir and Brady interest point detector[6]. Recently a comparison between local features is done by [7].

Many types of classifiers are used in action recognition context including discriminative classifiers like SVM, K-NN, LPBoost or generative classifiers like pLSA[1],

[1] Probabilistic Latent Semantic Analysis

LDA[2] and other topic models. Some advantages and disadvantage of discriminative and generative approaches is listed in Table 2. In this thesis, we use SVM and K-NN for evaluation of proposed methods.

Table 2: comparison of different classifiers.

Model	Accuracy	Number of train samples	Learn great number of classes	Using prior knowledge directly	Incremental learning (ability to increment number of classes)
Generative	more	fewer	Yes	Yes	Yes
Discriminative	less	more	No	No	No

Proposed methods

This thesis suggests 3 methods to improve accuracy of recognition. In first method, we used weighted features. pLSA [8] output is used to weight to each feature. pLSA uses EM algorithm for maximize an objective function and through it, two distribution function are estimated: distribution of each word over a class and distribution of each class over a video. We use them to compute probability of each word over a video:

$$P(w \mid d) = \sum_{z=1}^{Z} P(w \mid z) P(z \mid d)$$

And then weight of each feature for a specific action category is computed:

$$P(w \mid c_i) = \sum_{d_j \in c_i} P(w \mid d_j)$$

Figure 1 illustrated this process. It can be supposed as a hybrid classifier because the output of a generative classifier is used as inputs for discriminative classifier.

In second method, we extended pyramid spatial matching [9] and constructed spatio-temporal pyramids and classified actions by both constructing pre-computed kernels (using intersection operator) and χ^2 distance.

[2] Latent Dirichlet Allocated

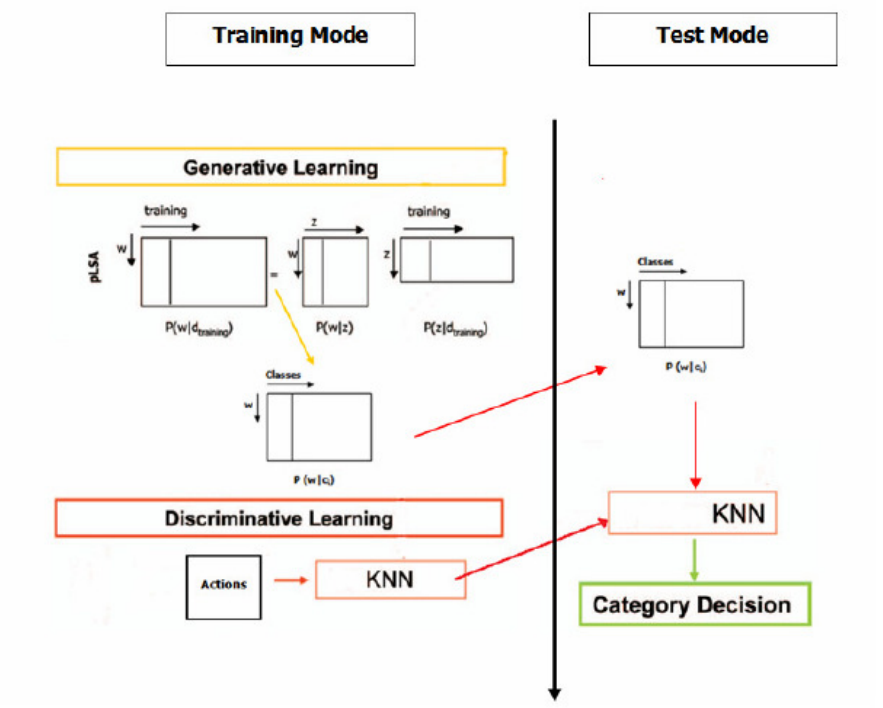

Figure 1: diagram of proposed weighting method.

In third method, we design another representation of action. In this model, each feature will be surrounded by a cube with specific height, length and width (Figure 2) and a local histogram is computed over it. Using this strategy, an adjacency matrix will be computed which is used as behavior vector (after flattening of 2D matrix to a vector).

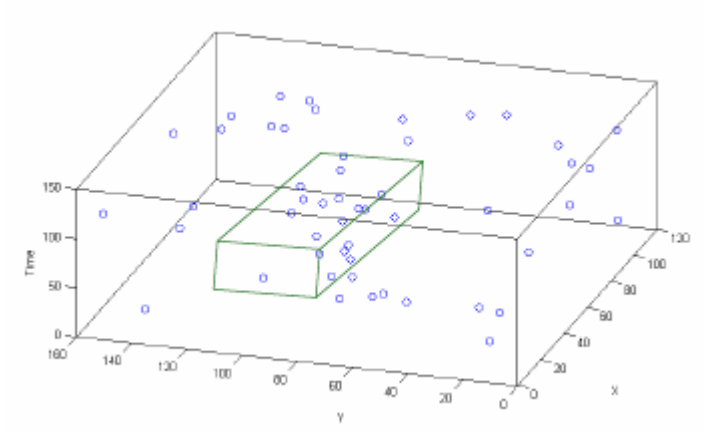

Figure 2: each feature enclosed by a cube.

Evaluations and experiment results

We evaluated our proposed methods on KTH and Weizmann datasets. We use k-fold cross validations and LOO[3] strategy to verify our results.

[3] Leave One Out

In this thesis, we use cuboids as space time features (Figure 3) and bag of words model (histogram of words over action) to describe behaviors. We generated words for each of action classes separately. Experiment shows that this kind of making dictionary, discriminate action's histograms more than overall clustering. However BoG model do not cooperate geometrical information of actions in order to recognition.

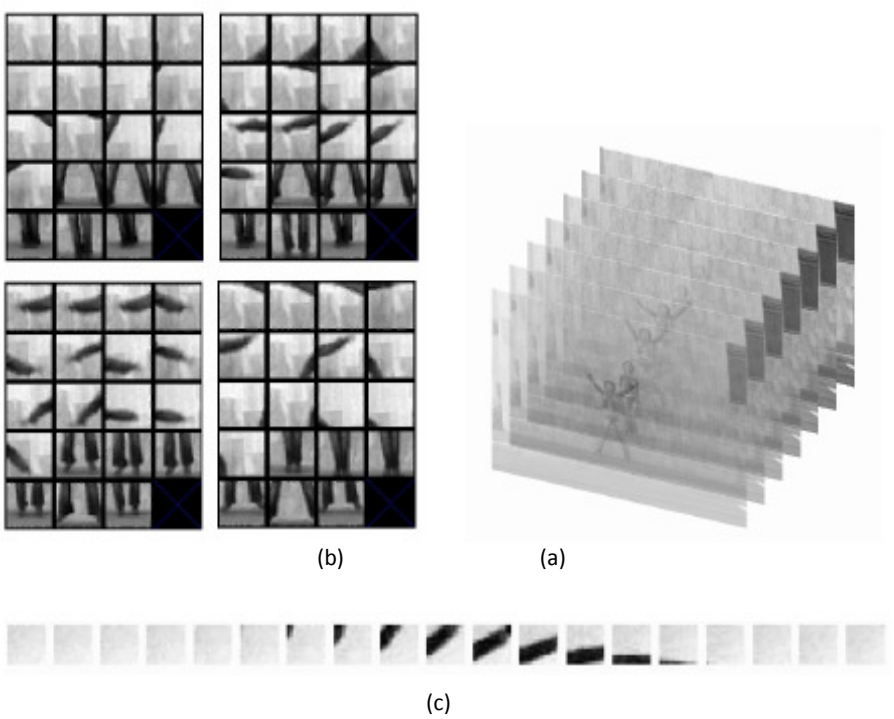

Figure 3: cuboids: a) an action, b) 4 frames of 19 detected features, c) one feature over time.

At first, we compared many descriptors of features including flattening values; global Histogramming, local Histogramming and 3D sift. During this experiment, we found that flattening gradients of x, y, t directions best fit to our frameworks as shown in figure 4.

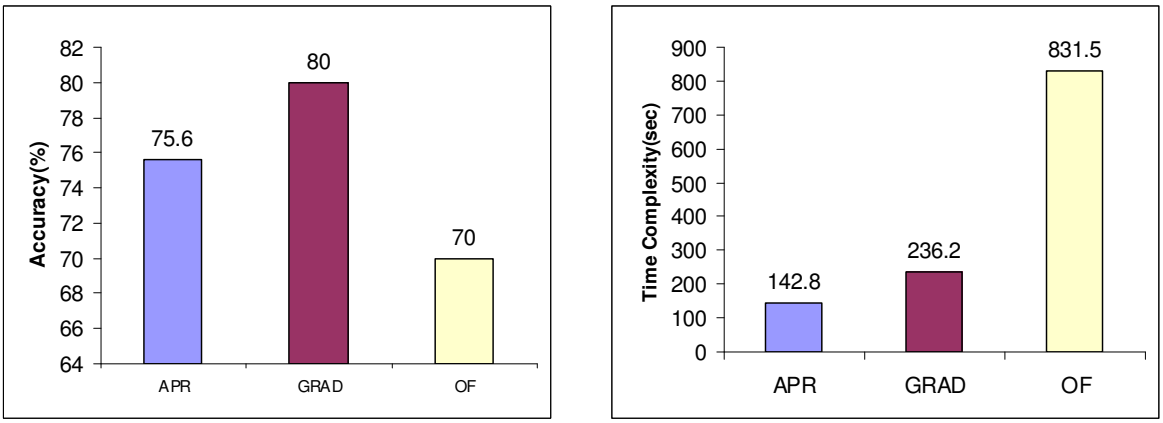

Figure 4: comparison of descriptors (accuracy and time complexity) on Weizmann dataset.

Figure 5 shows effect of choosing number of components in PCA in overall accuracy in order to dimension reduction.

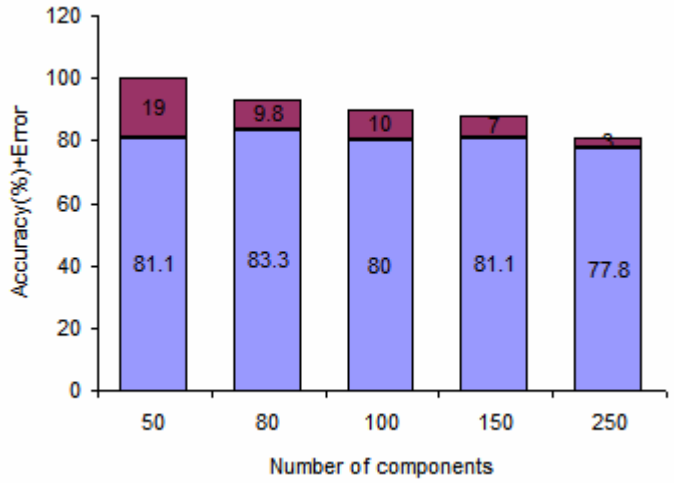

Figure 5: effect of PCA components. Blue error shows accuracy and red bar shows error.

Table 3 shows a comparison between methods applied on KTH dataset and our weighting method.

Table 3: accuracy of methods on KTH dataset.

Multiple action	Learning strategy	Recognition accuracy	method
×	Supervised	SVM	[9]
		71%	
✓	Unsupervised	pLSA	[10]
		83%	
×	Supervised	KNN	[3]
		81%	
×	Supervised	boosting	[4]
		63%	
×	Supervised	LPBoost	[11]
		89%	
×	Supervised	Semi-LDA	[12]
		91.2%	
×	Unsupervised	Semi-CTM	[12]
		90.3%	
×	Unsupervised	VWC Correlation	[13]
		94.2%	
×	Supervised	WX-SVM	[14]
		91.6%	
×	Supervised	Bio-Inspired	[15]
		91.7%	
×	Supervised	SVM	[11]
		87.4%	
×	Supervised	SVM	Proposed method
		92%	

Pyramid spatio-temporal matching is sensitive to shift and it does not show surprising results due to unsegmented data in time domain in KTH set. However it

just works for un-shifted data both in spatial and temporal domain.

Figure 6 shows results when 2D matrix descriptor is used. In this configuration, width and height (spatial domain) of cube is set to frame size and length is a variable in figure.

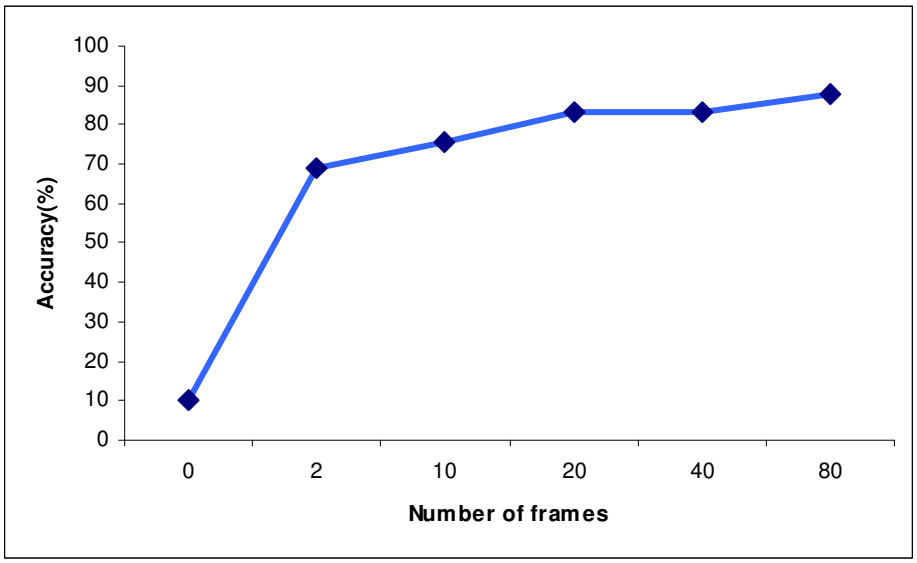

Figure 6: accuracy of proposed method 3, as number of frames (depth of cube) increases.

Future works:

Current actions are mostly performed in controlled setting for example without background motion, few clutter, and view-dependent. Evaluation on such data does not help much to discover true limitations of each method. However it is my opinion that evaluation of methods should be migrated to realistic scenes gradually. Working on true sport recordings, movies, and video data from the internet, will help us to discover the real requirements for action recognition, and it will help us to shift focus to other important issues involved in action recognition, such as segmentation of continuous actions, dealing with unknown motions, composite actions, multiple persons, and view invariance, for instance.

Also, a comparison between approaches mentioned in related works, in realistic actions, should be done. It helps us to find how much dynamics (information in time dimension) are important for recognition and is it necessary to interfere them explicitly (for example using HMM[4] or CRF[5]) or implicitly (for example using silhouettes, contours and local features).

Another challenging problem for action recognition is camera movements. While in most videos, camera is moving, a process to finding region of interest (action for this task) and detecting robust spatio temporal features is necessary and it is one of the most important tasks in order to recognition of actions in uncontrolled settings.

[4] Hidden Markov Models
[5] Conditional Random Fields

References

[1] I. Laptev, M. Marszałek, C. Schmid and B. Rozenfeld; "Learning realistic human actions from movies," *in Proc. CVPR'08*, Anchorage, US, 2008.

[2] I. Laptev, "Local Spatio-Temporal Image Features for Motion Interpretation," Ph.D Thesis, Computational Vision and Active Perception Laboratory, KTH, Stockholm, 2004.

[3] P. Dollar, V. Rabaud, G. Cottrell and S. Belongie, "Behavior recognition via sparse spatiotemporal features," In Proceedings of 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65-72, 2005.

[4] Y. ke, R. Sukthankar and M. Hebert, "Efficient visual event detection using volumetric features," In Proceedings of Tenth IEEE International Conference on Computer Vision, Vol. 161, pp. 166-173, 2005.

[5] A. Oikonomopoulos, I. Patras and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions," Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 36, no. 3, pp. 710-719, 2005.

[6] T. Kadir, A. Zisserman and M. Brady, "An affine invariant salient region detector," *In Proceedings of 8th European Conference on Computer Vision*, pp. 83-105, 2004.

[7] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. "Evaluation of local spatio-temporal features for action recognition," In BMVC, 2009.

[8] T. Hofmann, "Probabilistic latent semantic indexing," *In Proceedings of 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50-57, 1999.

[9] C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: a local SVM approach," In proceedings of 17th International Conference on Pattern Recognition, Vol.33, pp. 32-36, 2004.

[10] J.C. Niebles, H. Wang and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *Int. J. Comput. Vision*, vol. 79, no. 3, pp. 299-318, 2008.

[11] S. Nowozin, G. Bakir and K. Tsuda, "Discriminative Subsequence Mining for Action Classification," In Proceedings of 11th International Conference on Computer Vision ,pp. 1-8, 2007.

[12] W. Yang and G. Mori, "Human Action Recognition by Semilatent Topic Models," Pattern Analysis and Machine Intelligence, vol. 31, no. 10, pp. 1762-1774, 2009.

[13] J. Liu and M. Shah, "Learning Human Actions via Information Maximization," In Proceedings of IEEE conference on Computer Vision and Pattern Recognition, 2008.

[14] W. Shu-Fai, K. Tae-Kyun and R. Cipolla, "Learning Motion Categories using both Semantic and Structural Information," In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-6, 2007.

[15] H. Jhuang, T. Serre, L. Wolf and T. Poggio, "A Biologically Inspired System for Action Recognition," In Proceedings of IEEE 11th International Conference on Computer Vision, pp. 1-8, 2007.