Weakly Supervised Action Recognition using Implicit Shape Models

Tuan Hue Thi[1], Li Cheng[2], Jian Zhang[1], Li Wang[3], Shinichi Satoh[4]
[1]*National ICT of Australia & University of New South Wales, NSW, Australia*
[2]*Toyota Technological Institute at Chicago, Illinois, USA*
[3]*Southeast University, Nanjing, China*
[4]*National Institute of Informatics, Tokyo, Japan*

Abstract

In this paper, we present a robust framework for action recognition in video, that is able to perform competitively against the state-of-the-art methods, yet does not rely on sophisticated background subtraction preprocess to remove background features. In particular, we extend the Implicit Shape Modeling (ISM) of [10] for object recognition to 3D to integrate local spatio-temporal features, which are produced by a weakly supervised Bayesian kernel filter. Experiments on benchmark datasets (including KTH [11] and Weizmann [5]) verifies the effectiveness of our approach.

1. Introduction

Visual action recognition is a crucial problem in video analysis and understanding. It is nevertheless a challenge task due to the non-rigid object and motion shapes, variations due to changes in viewing angles and distances, and is further complicated by camera motion as well as background clutters. These difficulties prohibit practical attempts toward building a rigorous global model for each action class, as they often bear limited capacities to capture non-rigid shapes with varying poses, hence provides very little generalization for unknown data. Recent work such as [9, 11] partially address these issues by utilizing local features that are invariant to pose changes. On the other hand, to obtain satisfactory recognition rate, a de facto procedure is to apply dedicate preprocess to each of the video sequences using sophisticated background subtraction techniques, in order to extract accurate foreground objects [4, 5, 6, 7]. This procedure often involves heavy manual interactions and does not generalize well to novel videos.

In this paper, we propose a robust approach that is capable of addressing both limitations. Start with local features invariant to view and scale changes, our approach further applies an improved variant of the weakly-supervised Bayesian Learning work of Carbonetto *et al.* [2, 8] in Object Detection to videos, to focus on foreground actions with very little supervision. Moreover, we extend the Implicit Shape Model of Leibe *et al.* [10] to 3D. This enable us to robustly integrate the set of local features into a global configuration, while still being able to capture local saliency. Empirical experiments convincingly demonstrate the competitiveness of our proposed approach when comparing with the best known results.

2. Local Features as Video Representation

A video shot in our perspective is a complex set of local features under various configurations. Tackling action recognition this way as we discussed earlier helps to lighten the dependency on view and scale variance of action visual appearance. We adopt the existing Space Time Interest Point (STIP) detection technique from Laptev *et al.* [9] to detect points with high motion change. In addition, by observing that the certain regions around these detected points are also contributive to the action context, we refine STIP detection results with a post Inpainting procedure, which idea is similar to Image Inpainting described in [3] by Criminisi *et al.* The inpainting process starts on the boundary of connected STIP point regions, base on the median scale and frequency of these STIP points to generate hypothesis about whether other points in the neighborhood should be included. Figure 2 illustrates the effect of our improved technique, inpainted Space Time Interest Point (iSTIP), over the traditional STIP.

The surrounding areas of detected pixels are then described using a concatenation of Histogram of Oriented Gradients (HOG) and Histogram of Oriented Flow (HOF) [9]. In order to better organize the interest points in terms of their appearance, we use the ag-

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.858

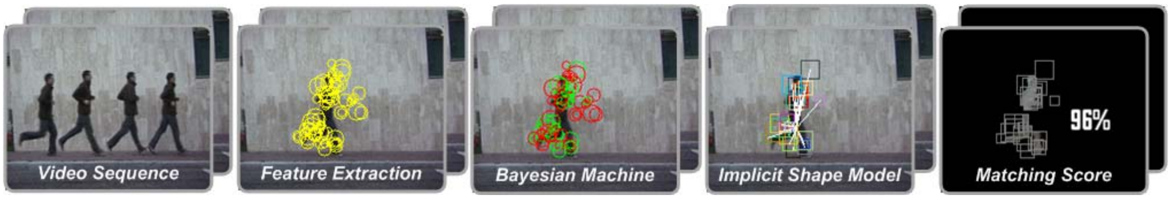
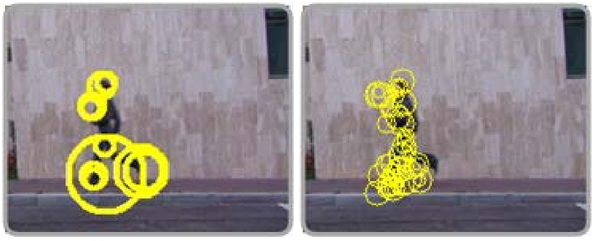


Figure 1. Framework Overview of Human Action Recognition System



(a) STIP Detection (b) iSTIP Detection

Figure 2. Feature Detection Improvement

glomerative clustering scheme from Agarwal and Roth [1] to group similar interest regions based on pairwise Normalized Greyscale Correlation (NGC) values, which helps to yield the highest similarity compactness of pixel appearance regardless of the number of clusters being produced. With this analysis, a video shot ν is now represented as a sparse set of all interest points $x_i(\iota, c)$ having cluster identity ι and 3D coordinate c.

3. Sparse Bayesian Learning

Among all detected interest points from the video shots, there are usually motion noise from the scattered background that do not contribute to the action motion. In fact, those points normally make the modeling computation much harder and in some cases might completely distract the core parts of the action. In order to filter out these irrelevant elements, we develop an extended version of the Sparse Bayesian Kernel Machine from Object Recognition work of Carbonetto *et al.* [2]. For each interest point $x_i(\iota, c)$ as notated in previous section, there will be associated a class label $y_i^k \in \{-1, 1\}$. The idea is to build a hierarchical Bayesian classifier model with parameters learned from the limited amount of available training data. Following [2], we adopt a sparse kernel machine for classification purpose, with the function between the posterior probability p and probit link Φ:

$$p(y_i = 1|x_i, \beta, \gamma) = \Phi(f(x_i, \beta, \gamma)) \quad (1)$$

with f is the regression function

$$f(x_i, \beta, \gamma) = \sum_{k=1}^{N} \beta_k \gamma_k \psi(x_i, x_k) \quad (2)$$

and $\psi(x_i, x_k) = exp(-(x_i - xk)/\sigma)$, the Gaussian kernel function of x_i with N feature points in the sampling. The two parameters of this classification model and the regression coefficients $\beta \triangleq [\beta_1\beta_2...\beta_N]$ and the feature selection vector $\gamma \triangleq [\gamma_1\gamma_2...\gamma_N]$, $\gamma_k \in \{0, 1\}$, implying the sparsity of this classification [8]. The discriminative classification now becomes calculation of the probability of a new point x' based on training data $\{x, y_k\}$, and model parameters $\theta = \{\beta, \gamma\}$

$$p(y'|x', x, y_k) = \int p(y'|x', \theta)p(\theta|x, y^k)d\theta \quad (3)$$

Figure 3 shows our successful adoption of this Sparse Bayesian Machine for the task of feature labeling on Weizmann [5] dataset, all red colored circles represent points noisy background while green circles are those positively labeled as parts of the action configuration, and denoted as *action elements* in our system.

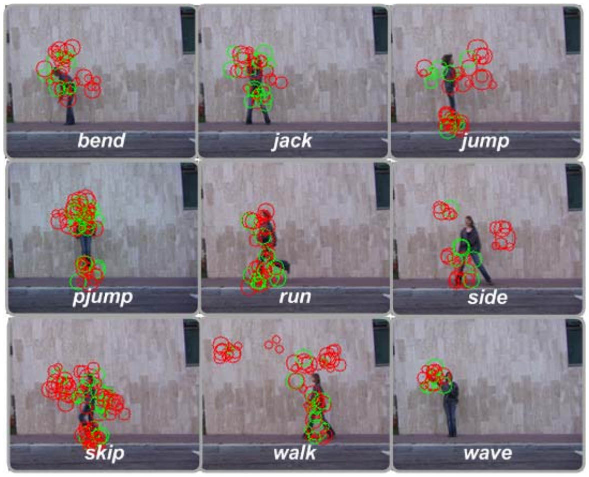


Figure 3. Feature Filter on Weizmann

4. Nonparametric Implicit Shape Model

The main part of our Action Recognition framework is the Implicit Shape Model, designed to integrate local attributes and global configuration of detected *action elements*). The idea is motivated by the Implicit Shape Model (ISM) approaches in the Object Categorization field. We come up with a flexible modeling technique that projects all *action element* properties onto an action hyperspace that consists of their spacetime coordinates c and their cluster identification ι. In this approach, by decomposing an action video into temporal slices or key frames, a *local action center* is considered as the centroid of all *action elements* in each slice, and a conceptual *global action center* is constructed as the mean center point along the projected trajectory of those *local action centers*. With these two notations, an *action element* can then be projected along with its relative local and global *action centers*. Formally, for each *action center* hypothesis at position x, we can factor the marginalization probability $p(x|e, c)$ for image patch e and coordinate system c based on observed cluster identities I_i according to [10]

$$p(x|e,c) = \sum_i p(x|e, I_i, c)p(I_i|e,c) \qquad (4)$$

$$= \sum_i p(x|I_i, c)p(I_i|e) \qquad (5)$$

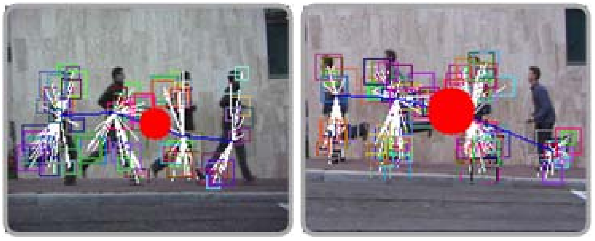
The calculation is done in a similar fashion to other Generalized Hough Transform models with vote casting from all elements for the most possible *action center*, and a 3D mean-shift search over the voting space is applied as a Parzen window probability density estimation to calculate the most voted action center volume, which in other words, the quantitative matching score of an unknown motion compared to the action model.

$$score(x) = \sum_k \sum_{x_j \in X} p(x_j|e_k, c_k) \qquad (6)$$

Figure 4 illustrates this voting procedure, where rectangles are the *action elements*, their colors represent visual cluster identities, *local action centers* are where the white lines converge and the big red circle is the nominated search space for *global action centers*.

5. Experimental Results

We run action classification on KTH [11] and Weizmann [5] dataset to evaluate our system performance. The KTH dataset contains nearly 2400 video shots of 6 different action classes, performed by 25 people under 4 different contexts. The Weizmann dataset has 92 video shots of 10 action classes, performed by 9 people. The detailed step-by-step results are shown in Figure 7, and the most representative *action elements* obtained from 16 actions of both datasets are displayed in Figure 5.

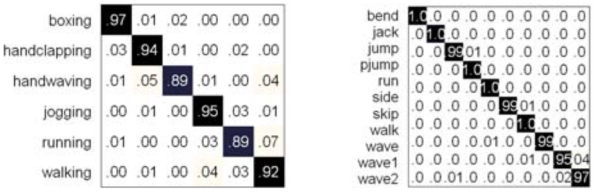


(a) *run* **ido** on Weizmann (b) *skip* **eli** on Weizmann

Figure 4. Voting Space for Action Center

Figure 5. Representative *action elements*

In order to produce a fair comparison with reported fully-supervised action recognition systems, we allow our training source to start with very little supervision and get increased after each round to reach the benchmark train/test amount of these two datasets, that is, *2/3 Split* on KTH, and *leave-one-out* on Weizmann. Figure 6 shows our classification confusion matrices obtained with maximum amount of training on each dataset.



boxing	.97	.01	.02	.00	.00	.00
handclapping	.03	.94	.01	.00	.02	.00
handwaving	.01	.05	.89	.01	.00	.04
jogging	.00	.01	.00	.95	.03	.01
running	.01	.00	.00	.03	.89	.07
walking	.00	.01	.00	.04	.03	.92

(a) KTH 92.67% (b) Weizmann 98.9%

Figure 6. Action Classification Results

We also conduct a comparison survey shown in Table 5 with average classification accuracy from our approach compared to the best reported works on these two datasets, given the same amount of training. Interestingly, while other works seem to work well for either one of the datasets (Fathi and Mori's got top for Weizmann but bottom for KTH, Grundmann *et al.* has best

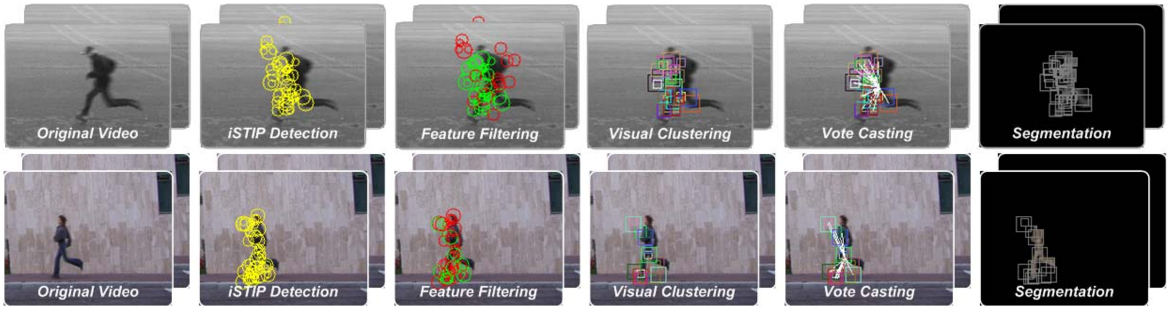


Figure 7. Classification procedure for action *run* from KTH and Weizmann dataset

score on KTH but lowest result on Weizmann), our system performs equally well on both datasets, having the second best classification score on KTH and Weizmann.

Author	KTH	Weizmann
Ours	***92.67***	***98.9***
Fathi and Mori [4]	90.50	100
Jhuang *et al.* [7]	91.70	98.8
Wang and Mori [12]	91.17	98.33
Grundmann *et al.* [6]	93.52	96.39

Table 1. Classification Average Accuracy

In addition, we carry out a thorough analysis on the robustness of our system toward supervision amount, to see how training amount affects on the overall classification accuracy. The two plots in Figure 8 prove that our system is able to pick up unknown data very quickly and once certain training amount is learnt, as little as 31.25% on KTH (corresponding to 89% classification accuracy) and 37.5% on Weizmann (with 93% accuracy), there is essentially no need to feed more training data. This demonstrates the beauty of nonparametric nature integrated in our Implicit Shape Model, highly generative even with small training amount.

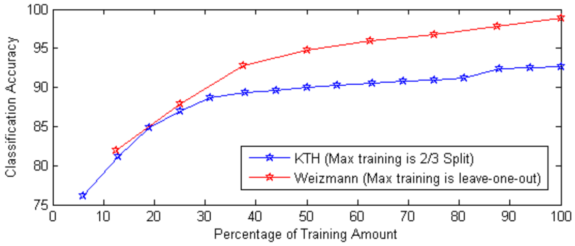


Figure 8. Accuracy vs Training Amount

6. Conclusion

In this paper we presented a new solution for the challenge of action recognition in video. Relying on two main components, Sparse Baysesian Machine and Implicit Shape Model, our system has successfully integrated the rich local feature attributes with the complex action global structure into one compact probabilistic model. Results on standard benchmarks have also demonstrated our approach can work under little supervision and still be highly abundant for regression.

References

[1] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *ECCV*, 2002.

[2] P. Carbonetto, G. Dorko, C. Schmid, H. Kuck, and N. de Freitas. Learning to recognize objects with little supervision. *IJCV*, 77(1-3):219–237, May 2008.

[3] A. Criminisi, P. Prez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. IP*, 13:1200–1212, 2004.

[4] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *CVPR*, pages 1–8, 2008.

[5] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *PAMI*, 2007.

[6] M. Grundmann, F. Meier, and I. Essa. 3d shape context and distance transform for action recognition. In *ICPR*.

[7] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*.

[8] H. Kuck, P. Carbonetto, and N. de Freitas. A constrained semi-supervised learning approach to data association. In *ECCV*, pages Vol III: 1–12, 2004.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic actions from movies. In *CVPR*.

[10] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV*, pages 17–32, 2004.

[11] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, 2004.

[12] Y. Wang and G. Mori. Human action recognition by semilatent topic models. *PAMI*, 31(10), 2009.