Action Recognition Using Spatial and Temporal Features with Kernel SVM

N.Nivetha

Assistant professor, Dept of Electronics and Communication Engineering
Unnamalai Institute of Technology, Tamil Nadu , India

Abstract

A new low-level visual feature, called Spatio-temporal context distribution feature of interest points is used to describe human actions. Each action video is expressed as a set of relative XYT coordinates between interest points listed pair wise in a local region. From the input image frames the Locally Weighted Word Context (LWWC) descriptor encodes the spatial context interest points rather than being limited to a single interest point and the Graph Regularized Nonnegative Matrix Factorization (GNMF) is used to encode the geometrical information by constructing a nearest neighbour graph. By extracting the kernel weights of the obtained feature variables , the kernel weighted SVM is modelled to jointly capture the compatibility between multilevel action features and action classes and the compatibility between multilevel scene features and scene classes. The contextual relationship between action classes and scene classes is derived using the kernel weight as a variable.

Keywords: **GMM, semantic correlations, SVM, kernel weight, action class, variable.**

1. INTRODUCTION

Human activity recognition is an important area of computer vision research and applications. The goal of the activity recognition is an automated analysis of ongoing events and their context from video data. Its applications include surveillance systems, patient monitoring systems, and a variety of systems that involve interactions between persons and electronic devices such as human-computer interfaces. Most of these applications require recognition of high-level activities, often composed of multiple simple actions of persons. Human activities are categorized into human actions, human-human interactions, human-object interactions, and group activities. Hierarchical state-based approaches and syntactic approaches interpret videos in terms of stochastic strings. Description-based approaches that analyze videos by maintaining their knowledge on activities' temporal, spatial, and logical structures.

1.1 Spatio-Temporal Features:

Spatio-temporal image processing involves an extra dimension of information. The spatial ones which is fore ground information extracted from frames and the temporal ones which is the background information. The temporal information, usually addressed in the context of motion detection, can provide extra cues about the contents, structure, and other high or low level information present in a scene.

2. LITERATURE REVIEW

The system [1] propose a new low-level visual feature, called spatio-temporal context distribution feature of interest points, to describe human actions. Each action video is expressed as a set of relative XYT coordinates between pairwise interest points in a local region. A novel mid-level class correlation feature to capture the semantic correlations between different action classes. Each input action video is represented by a set of decision values obtained from the pre-learned classifiers of all the action classes, with each decision value measuring the likelihood that the input video belongs to the corresponding action class. By treating the scene class label as a latent variable, we propose to use the latent structural SVM (LSSVM) model to jointly capture the compatibility between multilevel action features and action classes, the compatibility between multilevel scene features and scene classes, and the contextual relationship between action classes and scene classes

The system proposed in [2] explains a visual event recognition framework for consumer videos by leveraging a large amount of loosely labelled web videos .Observing that consumer videos generally contain large intra class variations within the same type of events, we first propose a new method, called Aligned Space-Time Pyramid Matching (ASTPM), to measure the distance between any two video clips. Second, we propose a new transfer learning method, referred to as Adaptive Multiple Kernel Learning ,fuse the information from multiple pyramid levels and features (i.e., space-time features and static SIFT features) and cope with the considerable variation in feature distributions between videos from two domains For each pyramid level and each type of local features, we first train a set of SVM classifiers based on the combined training set from two domains by using multiple base kernels from different kernel types and parameters, which are then fused with equal weights to obtain a pre learned average classifier.

This system proposed in [3] presents Action Bank, a new high-level representation of video. Action bank is comprised of many individual action detectors sampled broadly in semantic space as well as viewpoint space. This representation is constructed to be semantically rich and even when paired with simple linear SVM classifiers is capable of highly discriminative performance. This action bank is tested on four major activity recognition benchmarks. In all cases, performance of this method is better than the state of the art. The classifiers find strong transfer of semantics from the constituent action detectors to the bank classifier.

The system [4] proposed a novel approach based on Linear Dynamic Systems (LDSs) for action recognition. It introduce LDSs to action recognition. LDSs describe the dynamic texture which exhibits certain stationary properties in time. They are adopted to model the spatiotemporal patches which are extracted from the video sequence, because the spatiotemporal patch is more analogous to a linear time invariant system than the video sequence. The kernel principal angle to measure the similarity between LDSs, and then the multiclass spectral clustering is used to generate the codebook for the bag of features representation. A supervised codebook pruning method is used to preserve the discriminative visual words and suppress the noise in each action class. The visual words which maximize the inter-class distance and minimize the intra-class distance are selected for classification.

This system [5] proposed a new method human actions can be identified not only by the singular observation of the human body in motion, but also properties of the surrounding scene and the related objects. In this paper, we look into this problem and propose an approach for human action recognition that integrates multiple feature channels from several entities such as objects, scenes and people. We formulate the problem in a multiple instance learning (MIL) framework, based on multiple feature channels. By using a discriminative approach, we join multiple feature channels embedded to the MIL space.

3. PROPOSED SYSTEM

This method aims to recognize the actions of one or more subjects from a series of observations on the subjects actions and the background conditions. The Locally Weighted Word Context (LWWC) descriptor encodes the spatial context interest points rather than being limited to a single interest point. Graph Regularized Nonnegative Matrix Factorization (GNMF) is used to encode the geometrical information by constructing a nearest neighbor graph

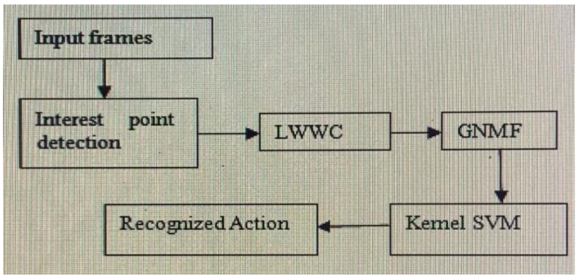


Fig 3.1- Block diagram of the proposed system

4. IMPLEMENTATION AND DESCRIPTION

4.1Input Frames

Videos from UCF 50 sports database are converted into frames and 10 frames are selected from 5 different videos. The features are sampled from all 10 frames

4.2 Interest Point Detection

From the input image frames interest points are detected using LWWC

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$$

I = I(x,y,t) with x & y as pixel dimensions and
t denotes the stack of images
g(x:y:σ)- 2D gaussian smoothing kernel
h_{ev} , h_{od} – quadrature pair of 1D gabor filter denoting even and odd parameters

4.2.1Gabor filter parameters

Even parameter

$$h_{ev}(t;\tau;\omega) = -\cos(2\pi tw)e^{-t^2/\tau^2}$$

Odd parameter

$$h_{od}(t;\tau;\omega) = -\sin(2\pi tw)e^{-t^2/\tau^2}$$

4.3 LOCALLY WEIGHTED WORD CONTEXT DESCRIPTOR (LWWC)

A context-aware descriptor called locally weighted word context (LWWC) is the low-level descriptor. LWWC encodes spatial context information rather than being limited to a single interest point. Such spatial context information is extracted from neighbouring of interest points, and can be used to improve the robustness and discriminability of the descriptor.

4.4 GNMF(graph regularized nonnegative matrix factorization

To extract the high level action unit

Let , i=1,…..C and j=1, …,n_i denote the d-dimensional low level feature representation of the j-th video in class.

The representation in class I form matrix

$$Y^i = [y_i^1 y_i^{ni}] \in R^{d*ni}$$

GNMF minimizes the objective function

$$\| Y^i - UV^T \|_F^2 + \lambda Trace(V^T LV)$$

$U \in R^{d*si}$ and $V \in R^{ni*si}$ are two non negative matrices.

- L=D-W called graph laplacian
- W-symmetric and nonnegative similarity matrix.
- D is the diagonal matrix whose entries are column sums of W.
- Matrix U – column vectors of U was defined by the action units belonging to action class i.
- Matrix V^T – each column of V^T is a low dimensional representation of corresponding column of Y^i with respect to the new bases.

4.4.1 GNMF-WEIGHING FUNCTION

Consider a graph with N vertices, for each data point xj with nearest neighbors of 2 nodes i and j. The weight matrix W on the graph can be calculated by any one of the methods.

1. 0-1 weighting : W_{jl} =1
2. Heat Kernel Weighting : $W_{jl}=e^{-\|x_j - x_l\|^2}/\sigma$
3. Dot-product Weighting : $W_{jl} = x_j^T x_l$

- For image data, the heat kernel weight may be the better choice .
- W_{jl} is only for measuring closeness , the different weighting schemes was not treated separately.
- The heat kernel weight is adopted as

$$\boxed{W_{jl} = 1/6\ ||y_i^j - y_i^l||^2}$$

4.5 KERNAL SVM

A limitation of LSVM is that they rely on linear models. For many computer vision tasks, linear models are suboptimal and nonlinear models learned with kernels typically perform much better. Kernel SVM (KSVM) – a new learning framework that combines latent SVMs and kernel methods. The use of kernel weights make non-separable problem separable and it Maps the data into better representational space.

The dual form of the discriminant function is defined as

$$\mathcal{D}(\alpha^*, \beta^*) = \max_{\alpha,\beta} \sum_i \alpha_i + \sum_j \sum_h \beta_{j,h} - \frac{1}{2} || \sum_i \alpha_i \phi(x_i, h_i) - \sum_j \sum_h \beta_{j,h} \phi(x_j, h) ||^2$$

The scoring function for the testing images X^{new} can be kernalized as follows,

$$\max_{h^{new}} \left(\sum_i \alpha_i^* k(\phi(x_i, h_i), \phi(x^{new}, h^{new})) - \sum_j \sum_h \beta_{j,h}^* k(\phi(x_j, h), \phi(x^{new}, h^{new})) \right).$$

4.5.1 Iterative Algorithm

- Fix α and β, compute the optimal $\{h_i\}^*$ by

$$\{h_i\}^* = \arg\max_{\{h_i\}} \frac{1}{2}\|\sum_i \alpha_i \phi(x_i, h_i) - \sum_j \sum_h \beta_{j,h} \phi(x_j, h)\|^2$$

- Fix $\{h_i\}$, compute the optimal (α^*, β^*) by

$$(\alpha^*, \beta^*) = \arg\max_{\alpha,\beta} \left\{ \sum_i \alpha_i + \sum_j \sum_h \beta_{j,h} - \frac{1}{2}\|\sum_i \alpha_i \phi(x_i, h_i) - \sum_j \sum_h \beta_{j,h} \phi(x_j, h)\|^2 \right\}$$

4.6 EXPERIMENTAL SETTING:

For interest points detection, the spatial and temporal scale parameters σ and τ are empirically set by σ = 2 and τ = 2.5, respectively. The size of cuboid is empirically fixed as 7×7×5 and 1000 interest points are extracted from each video. For the spatio-temporal (ST) context distribution feature of interest points, the number of space-time scales is fixed to five and the number of Gaussian components in GMM (i.e., K) is set to 2000.

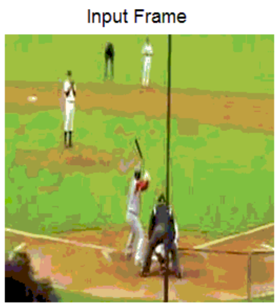
5. RESULTS AND DISCUSSION

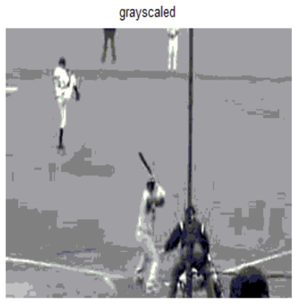
5.1 Videos selected

- Baseball
- Basketball
- Swimming
- High jump
- Golf

5.2 Baseball – Input Frame

5.3 Gray Scaled Frame

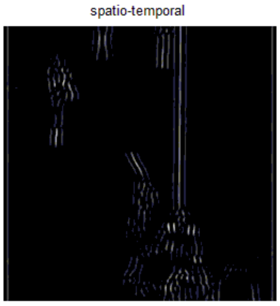




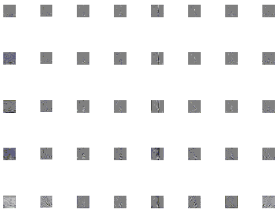
5.4 Resized Frame

5.5 Smoothened Frame

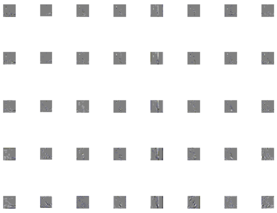
5.6 Spatio-Temporal Filtered



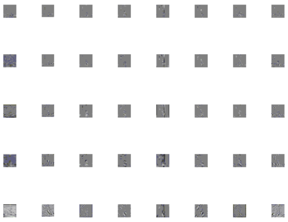
5.7 Real Parts Of Gabor Filter



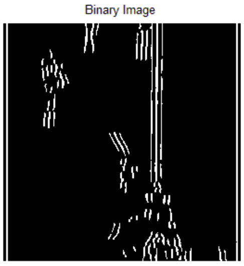
5.8 Imaginary Parts Of Gabor Filter



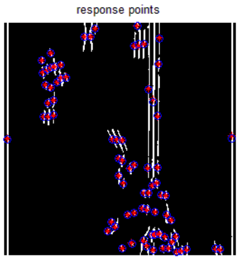
5.9 Magnitude Of Gabor Filter



5.10 Binary Scaled Image

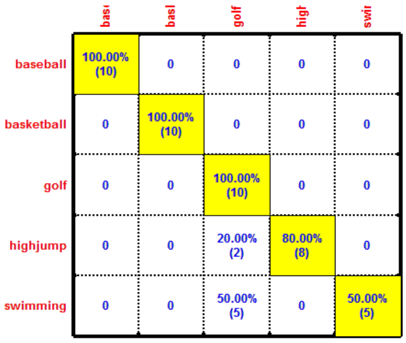


5.11 Response Points From Frame



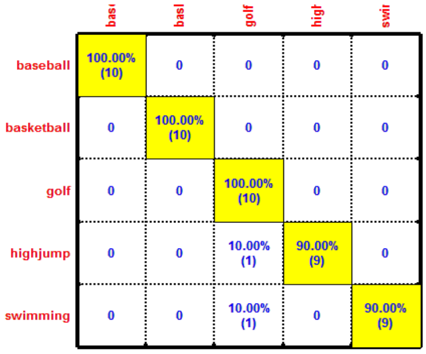
5.12 Confusion matrix:

5.12.1 without kernel weight



	basi	basl	golf	higt	swii
baseball	100.00% (10)	0	0	0	0
basketball	0	100.00% (10)	0	0	0
golf	0	0	100.00% (10)	0	0
highjump	0	0	20.00% (2)	80.00% (8)	0
swimming	0	0	50.00% (5)	0	50.00% (5)

5.12.2 with kernel weight



	basi	basl	golf	higt	swii
baseball	100.00% (10)	0	0	0	0
basketball	0	100.00% (10)	0	0	0
golf	0	0	100.00% (10)	0	0
highjump	0	0	10.00% (1)	90.00% (9)	0
swimming	0	0	10.00% (1)	0	90.00% (9)

5.13 COMPARISON OF RESULTS

S.No	Classifier	Without Kernel Weight	WithKernel Weight
1.	SVM (ST+HOG)	86%	96%

6. SUMMARY

The videos are converted into frames and the input frames are converted into grayscale images and then they are resized to 130*66 pixels for applying the filters. The Gaussian filtering and Gabor filters are applied to the resized images and the response points are obtained from each frames for all the action videos. The response point features are extracted into vector for matrix factorization and to the kernel weights are obtained for each action units. The obtained kernel weights has to be trained along with the action units in the kernel SVM classifier to classify the actions

7. CONCLUSION

An efficient feature extraction method is presented by using the Spatial and Temporal features from the video data. The foreground and background informations are obtained by applying the LWWM and Gabor filter parameters. Then the interest points are trained using video specific GNMF(graph regularized nonnegative matrix factorization. The GNMF is used to increase the accuracy in classification. The action and scene class models are used along with the interactions between them. The classification process is done using SVM, but for using action and scene class model makes it complicated. So the kernel weight is extracted from the SVM and used as a variable in classification to increase the accuracy.

8. REFERENCES

1. Xinxiao Wu, Dong Xu, Lixin Duan, Jiebo Luo, and Yunde Jia," Action Recognition Using Multilevel Features and Latent Structural SVM" IEEE transactions on circuits and systems for video technology, vol. 23, no. 8, Aug 2013.
2. X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 489–496.
3. Sadanand, S, Corso, JJ 2012, 'Action bank: A high-level representation of activity in video', in Proc. IEEE CVPR, pp. 1234 –1241.
4. N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," in Proc. ECCV, pp. 494–507, Sep. 2010.
5. Wang, H, Ullah, MM, Klaer, A, Laptev, I, Schmid, C 2009, 'Evaluation of local Spatio-Temporal features for action recognition', in Proc. BMVC, pp. 1–11.
6. Kovashka, A, Grauman, K 2010, 'Learning a hierarchy of discriminative space- time neighborhood features for human action recognition', in Proc. IEEE CVPR, pp. 2046–2053.
7. X. Wu, Y. Jia, and W. Liang, "Incremental discriminant - analysis of canonical correlations for action recognition," Pattern Recognition vol. 43, no. 12, pp. 4190–4197, Dec. 2010.