Action Recognition using High-Level Action Units

Paul. T. Jaba
M.E, Department of Computer Science
St. Joseph College of Engineering,
Sriperumbudur, India
write2jaba@gmail.com

Ms. Jackulin Asha G. S.
Department of CSE,
St. Joseph College of Engineering,
Sriperumbudur, India

***Abstract*—** Vision-based human recognition is the process of naming image sequences with action labels. In this project, a model is developed for human activity detection using high-level action units to represent human activity. Training phase learns the model for action units and action classifiers. Testing phase uses the learned model for action prediction.Three components are used to classify activities such as New spatial-temporal descriptor, Statistics of the context-aware descriptors, Suppress noise in the action units. Representing human activities by a set of intermediary concepts called action units which are automatically learned from the training data. At low-level, we have existing a locally weighted word context descriptor to progress the traditional interest-point-based representation. The proposed descriptor incorporates the neighborhood details effectively. At high-level, we have introduced the GNMF-based action units to bridge the semantic gap in activity representation. Moreover, we have proposed a new joint *l*2,1-norm based sparse model for action unit selection in a discriminative manner. Broad experiments have been passed out to authorize our claims and have confirmed our intuition that the action unit based representation is dangerous for modeling difficult activities from videos.

Keywords- *Action unit, sparse representation, nonnegative matrix factorization, action recognition*

*****

I. INTRODUCTION

HUMAN activity detection has a wide range of applications such as video content analysis, activity surveillance,and uman-computer interaction [2]. As one of the most active topics in computer vision, much work on human activity detection has been reported. In most of the traditional approaches for human activity detection, activity models are normally constructed from patterns of low-level features such as appearance patterns [4], optical flow, space-time templates, 2D shape matching, trajectory-based representation and bag-of-visual-words (BoVW). However, these features can hardly distinguish rich semantic arrangement in activity.

Inspired by latest growth in object categorization, introducing a high-level concept named "action unit" to describe human actions.

For sample, the "golf-swinging" activity contains some representative motions, such as "arm swing" and "torso twist". They are hardly described using the low-level features mentioned above. Otherwise, some connected space-time interest points, when joint together, we can characterize a representative action. In addition, the key frame is essential to describe an activity; and a key frame may be characterize by the co-occurrence of space-time interest points extracted from the frame. The representative actions and key frames both mirror some action units, which can then be used to represent action classes. With the above examination, propose using high-level action units for human activity representation. Usually, from an input human activity video, hundreds of interest points are first extracted and then agglomerated into tens of action units, which then compactly represent the video. Such a representation is more discriminative than traditional BoVW model. To utilize it for activity detection, we address the following three major issues.

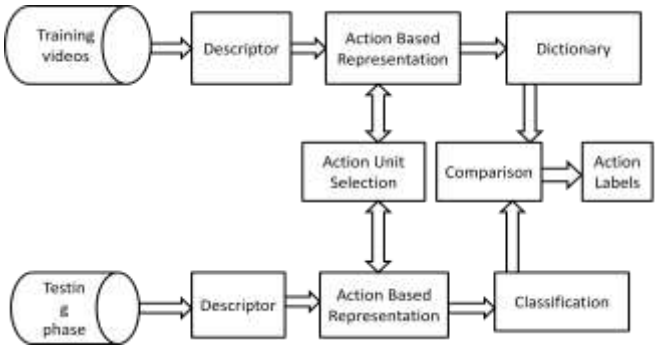

Fig 1:System Architecture

1. *Selecting low-level features for generate the action unit.* Some of the aforementioned features needs reliable tracking or body pose evaluation, which is hard to attain in practice. The interest-point-based representation avoid such requirement while being robust to noise, occlusion and geometric deviation. But conventional bag-of-visual-words models (BoVW) use only features from single interest points and avoid spatial-temporal context details. To address this issue, we propose a latest context-aware descriptor that incorporates context details from adjacent interest points. This way, the new descriptor is more discriminative and robust than the traditional BoVW.

2. Building an action unit set to represent all action classes under examination. Nonnegative Matrix Factorization (NMF) [5] has received considerable notice and has been exposed to capture part-based representation in the human brain as well as visualization tasks. We propose using graph regularized Nonnegative Matrix Factorization (GNMF) to encode the geometrical details by generating a nearest neighbor graph. It finds a part-based representation in which two data points are joined if they are satisfactorily close to each other. The GNMF-based action units are automatically learned from the training data and are capable of capturing the intra-class variation of each action class.

3. Choosing discriminative action units and suppress noise in action classes. We propose a new action unit selection method depends on a joint $l_{2,1}$-norm minimization. We first begin the $l_{2,1}$-norm for vectors. Sparse model based on such norm is robust to outliers and the regularization can lead selecting action units across intra-class samples. The dictionary learning process captures the fact that actions from the same class share similar action units. In this work we target learning high-level action units to characterize and sort human actions. For this intention we progress over the usual interest point feature and suggest an action unit based solution, which is further enhanced by an action unit selection procedure. In review, the training phase learns the method for action units and the action classifier on the activity unit-based representation. The testing phase uses the learned model for activity prediction.

II. RELATED WORK

Activity detection has been broadly explored in the computer vision community. In recent times, a little attempts have been made to utilize the mid- or high-level concepts for human activity detection. Mutual information maximization techniques to discover a compact mid-level codebook and utilize the spatial-temporal pyramid to exploit temporal information. Extract discriminative flow features within overlapping space-time cells and select mid-level features via AdaBoost. Unfortunately, the global binning makes the representation sensitive to place or time shifts in the clip segmentation, and using predetermined fixed-size spatial-temporal grid bins assumes that the proper volume scale is known and uniform across action classes. Such uniformity is not inherent in the features themselves, given the large differences between the spatial-temporal distributions of the features for different activities. Use the secret conditional random fields for activity detection. The authors model an action class as a basis template and a onstellation of secret "parts", where the secret "part" is a group of local patches that are implicitly correlated with some intermediate representation. Initially track the human with 3D pose estimation, which is then used for activity detection. Use diffusion maps to automatically learn a semantic visual vocabulary from abundant quantized mid level features, each one represented by the vector of pointwise mutual information. But the vocabularies are created for individual categories, thus they are not universal and general enough, which limits their applications. Learn data-driven attributes as the suppressed variables. The authors augment the manually-specified attributes with the automatically learned attributes to provide a complete characterization of human actions. Compared with traditional lowlevel features, it is obvious that human actions are more effectively represented by considering multiple highlevel semantic concepts. However, the current learning-based visual representations obtain labels from the entire video, and hence include background clutters which may degenerate learning effectiveness.

III. PROPOSED ALGORITHM

A. *Locally weighted word context descriptor*

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

These interest points are initially described by histograms-of-opticalflow (HOF) and histograms-of-oriented-gradients (HOG), which respectively characterize the motion and appearance within a volume surrounding an interest point. Afterwards, we employ the k-means algorithm on these features to create a vocabulary of size K. Following the BoVW, each interest point is then converted to a visual word. Finally, for each interest point together with its N − 1 nearest interest points, the Locally Weighted Word Context descriptor (LWWC) is calculated .

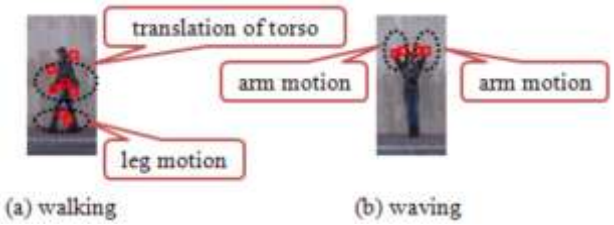

Fig 2: Action units

A. *GNMF-based action units*

NMF is a factorization algorithm for analyzing nonnegative matrices. The nonnegative constraints allow only additive combinations of different bases. This is the most significant difference between NMF and other matrix factorization methods such as singular vector decomposition

(SVD). NMF can learn a part-based representation. But it fails to discover the intrinsic geometrical and discriminative structure of the data space, which is essential to the real-world applications. If two data points are close in the intrinsic geometry of the data distribution, the representations of these two points with respect to the new bases should be still close to each other. GNMF aims to solve this problem.

The action-unit-based representation has two main advantages. First, it is compact since only tens of action units to describe an action video. This is more efficient in than BoVW models where hundreds of interest points are needed. Second, some low-level local features are not discriminative, and even have negative influence on classification. The process of learning class-specific action units can suppress such noises. The matrix factorization algorithm extracts the representative action units for each action class. The representative action units should exist in all the videos belonging to the same action class. Some low-level local features that only exist in a few intra-class videos are suppressed by the algorithm mentioned above, and are not used for constructing the highlevel action units. The learned class-specific action units can exhibit the characteristic of each action class. So, the proposed action-unit-based representation is more powerful for classification.

c) *Robust action unit selection based on joint l2,1-norm*

Typically, it approximates the input signal in terms of a sparse linear combination of the given over complete bases in dictionary. Such sparse representations are usually derived by linear programming as an l1-norm minimization problem. But the l1-norm based regularization is sensitive to outliers.

Inspired by the l2,1-norm of a matrix , first introduce the l2,1-norm of a vector. Moreover, we propose a new joint l2,1-norm based sparse model to select the representative action units for each action class. The proposed sparse model mainly has two advantages for classification based action unit selection:

- First, the l2,1-norm of the matrix in our sparse model encourages that the samples from the same action class are constructed by similar action units, and the action units which only appear in several intra-class samples are suppressed.

- Second, each action class has its own representative action units. The l2,1-norm of the vector in our sparse model encourages each sample is constructed by the action units from the same class.

IV. DISCUSSIONS

The main objective of the project is to develop a new method to represent and classify videos using high-level action units. To improve classification accuracy.

At low-level, presented a locally weighted word context descriptor to progress the traditional interest-point-based representation. The proposed descriptor incorporate the locality information effectively. At high-level, introduce the GNMF-based action units to connection the semantic gap in activity representation.

V. EXPERIMENTAL RESULTS

A. *Data Sets*

Five action data sets are used in our evaluation:

- The KTH action data set
- The UCF Sports data set
- The UT-Interaction data set
- The UCF YouTube data set
- The Hollywood2 data set

B. *Effects of the LWWC Descriptor*

Experiments are conducted to evaluate the influence of the neighborhood information in the LWWC descriptor. the recognition rates corresponding to different scales of neighborhood information covered by the proposed descriptor on the KTH and the challenging UT-Interaction data sets. Traditional interest point based methods only utilize features of a single interest point. It can only describe a very small area. So the accuracy can be easily influenced by noise.

The recognition rate is 93.99% on the KTH data set.

C. *Analysis of the Action Unit Selection*

To evaluate action unit selection, we compare the performances of the original GNMF-based action units with the one using action unit selection, In the KTH data set, the action unit selection significantly boosts the performance from 92.65% to 95.49%. On the UT-Interaction data set, it again significantly boost the recognition accuracies from 80.0% to 81.7% (on Set 1) and from 70.0% to 80.0% (on Set 2). These results clearly validate that the proposed joint l2,1-norm based action unit selection method is effective to improve the recognition performance.

D. *Experiments on the KTH Data Set*

The experimental results validate the effectiveness of the proposed method. Furthermore, we compare the performances of different baseline approaches (such as traditional single interest point feature, the proposed LWWC descriptor, l1-norm based sparse model, and our action unit selection approach),and study the contribution of each part in our method. The accuracy of traditional single interest point feature is 91.80%. The LWWC descriptor, the accuracy is 92.65%. When the action unit selection based on the traditional single interest point feature, the accuracy is 93.99%. Combining both, the accuracy reaches 95.49%. When

use the l1-norm based on LWWC descriptors, the accuracy is 92.82%. The study demonstrates that each of the proposed approaches offers more discriminative power than the BoVW baseline, and the l2,1-norm based action unit selection approach obtains better performance than the l1-norm based sparse model. It further validates the effectiveness of the high-level descriptor for classification. Our method, which combines the low-level LWWC descriptor with the high-level action unit selection, achieves the best performance.

E. ***Experiments on the UCF Sports Data Set***

The confusion matrix across all scenarios in the leave-one-out protocol on the UCF Sports data set. Our method works well on most actions. For example, the recognition accuracies for some actions are high up to 100%, such as "diving" and "lifting". There are complex backgrounds in this data set, and some actions are very similar and challenging for recognition, such as "golfing", "horseback riding", and "running". We conduct further experiments on the UCF Sports data set to study different components of the proposed approach, similar as on the KTH data set.

F. ***Experiments on the UT-Interaction Data Set***

The action videos in UT-Interaction are divided into two sets. To generate the codebook, empirically set the codebook size k to 500 in set 1, and set the codebook size k to 300 in set 2. We set the tradeoff parameters $\gamma 1 = \gamma 2 = 0.2$ for both sets.The leave-one-out test strategy is performed. The LWWC descriptor performs better than the action unit selection method in both sets, and the combination of them provides the best performance. The l2,1-norm based action selection method outperforms the l1-norm based sparse model.

G. ***Experiments on the UCF YouTube Data Set***

The confusion matrix across all scenarios on the UCF YouTube data set. In Fig. 12, we compare per action class performances of relative methods including cuboid, LWWC, action unit selection and some previously reported interest-point-based methods, and our method outperforms others. The overall mean accuracy of our method with the reported by previous researchers. Our average recognition accuracy is 82.2%, which is comparable to the state-of-the-art performance and outperforms other interestpoint- based methods.

H. ***Experiments on the Hollywood2 Data Set***

The performance of our method, and tests the contribution of each part in our method to the recognition accuracy respectively. The accuracy of traditional single interest point feature is 47.9%. If the utilization of the LWWC descriptor, then the accuracy is 50.1%. When we only adopt the action unit selection based on the traditional single interest point feature, the accuracy is 54.5%. In combination, the accuracy of our method is 56.8%. The overall mean accuracy of our method with the results reported by previous researchers. Our average recognition accuracy is better than or comparable to the state-of-the-art performances.

Fig 3: Object tracking using block matching algorithm

Fig 4: Bend action

Fig 5: Action unit

VI. CONCLUSION

In this paper, we have proposed to represent human activities by a set of intermediary concepts called action units which are automatically learned from the training data. At low level, presented a locally weighted word context descriptor to improve the traditional interest-point-based representation.The

proposed descriptor incorporates the neighborhood information effectively. At high level, we have introduced the GNMF-based action units to bridge the semantic gap in action representation. moreover, proposed a new joint $l_{2,1}$-norm based sparse model for action unit selection in a discriminative method. Extensive experiments have been carried out to validate our claims and have confirmed our intuition that the action unit based representation is critical for modeling complex activities from videos.

REFERENCES

[1] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2009,pp. 1778–1785.

[2] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Sep. 2008.

[3] H. Wang, C. Yuan, W. Hu, and C. Sun, "Supervised class-specific dictionary learning for sparse modeling in action recognition," *Pattern Recognit.*, vol. 45, no. 11, pp. 3902–3911, 2012.

[4] I. Kotsia, S. Zafeiriou, and I. Pitas, "Texture and shape information fusion for facial expression and facial action unit recognition," *Pattern Recognit.*, vol. 41, no. 3, pp. 833–851, 2008.

[5] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[6] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3337–3344.

[7] T. Berg, A. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," in *Proc. ECCV*, 2010,pp. 663–676.

[8] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal.Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.

[9] S. Sadanand and J. Corso, "Action bank: A high-level representation of activity in video," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2012, pp. 1234–1241.

[10] M. Raptis and L. Sigal, "Poselet key-framing: A model for human activity recognition," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2013, pp. 2650–2657.