International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Action Recognition Using Action Unit Based Feature Extraction and SURF Descriptor

Arya P Muraleedharan[1], Amrutha V Nair[2]

[1]M .Tech Student, Sree Buddha College of Engineering, Elavumthitta, Kerala, India

[2]Assistant Professor, Department of Electronics and Communication Engineering, Sree Buddha College of Engineering, Elavumthitta, Kerala, India

Abstract: ***A model is proposed for human action recognition using a new concept of action units to represent human actions in videos. Two stages are involved in our approach. First, training stage learns the model for action units and action classifiers. Second, testing stage uses the learned model for action recognition. Different interconnected components are used for action recognition. First involves the use of a new descriptor named SURF which is the fastest one with good performance and shows its advantages in rotation, blur and illumination changes. SURF uses square-shaped filters which forms an approximation of Gaussian smoothing. Second, involves learning action units using a factorizing algorithm called the graph regularized nonnegative matrix factorization, which helps to encode geometrical information. Third, a model is proposed to select the discriminative action units for better action recognition. SVM model is adopted as the predictive model.***

Keywords: Action unit, support vector machine, graph regularized non-negative matrix factorization, speeded up robust feature extraction, action recognition

1. Introduction

Activity recognition aims to recognize or analyze the human activities in videos. Recognizing human action has a wide range of applications such as video surveillance, human-computer interface, analysis of sports events, video indexing and browsing, recognition of gestures. Understanding the perception of actions in both humans and animals is an important area of research crossing the boundaries between several scientific disciplines from computer science to brain science. Activity recognition can be referred to as plan recognition, goal recognition, intent recognition, behavior recognition due to its many-faceted nature. Traditional approaches for action recognition such as 2-D shape matching, appearance patterns, bag-of-visual-words etc utilize certain low level features and ignore context information.

Here a new concept called action units are used to represent human actions in videos. Key frames helps to better describe an action. The key frames reflect some action units which can be used to represent some action classes. An input action video consists of hundreds of interest points which can be agglomerated into tens of action units, which compactly represent the entire video.

A new descriptor called Speeded Up Robust Feature descriptor is used to detect and describe features in images. Interesting points on the object can be extracted to provide a feature description for any object in an image. This feature description, obtained from a training image, can then be used to identify or detect the object when attempting to locate the object in a test image. It is important that the features extracted from the training image be detectable even under certain circumstances like changes in image scale, noise and illumination, to perform a reliable recognition. SURF descriptor helps to find such points in a fastest way with good performance.

Nonnegative Matrix Factorization is used for analyzing non-negative matrices. The resulting matrices is easier to inspect because of this non-negativity. Matrix factorization method is widely used in pattern recognition, information retrieval and computer vision. We propose a graph regularized Nonnegative Matrix Factorization which by constructing a nearest neighbor graph encodes the geometrical information. The action units are automatically learned from the training dataset and are capable of capturing the intra-class variation of each action class. A model based on the concept of sparsity is used to preserve the representative items and suppress noises. The discriminative dictionary learning is performed by the SVM model.

This paper introduces a novel method which combines the concept of action unit and graph regularized non-negative matrix factorization to detect actions in videos.

2. Literature Survey

Action recognition has been widely used in different applications and different concepts are used for recognizing human actions. Haoran Wang and Yaun et al. [1] use high-level action units to represent human actions in videos. A new context aware descriptor is used.. M. Blank, M. Irani, and R. Basri et al. [2] describes action in video sequences as silhouettes of a moving torso that undergoes motion. A method developed for the analysis of 2D shapes with volumetric space-time shapes induced by actions in videos is generalized. D. Cai, X. He, J. Han, and T. S. Huang et al. [3] discussed about Matrix factorization techniques that is widely used in information retrieval, computer vision, image segmentation and pattern recognition. Nonnegative Matrix Factorization (NMF) has received considerable attention due

to its psychological and physiological in representing actions as part based units in the human brain.

S. Ali, A. Basharat, and M. Shah et al. [4] uses the concepts of chaotic theory to analyze nonlinear dynamics of human actions. The representation of the non-linear dynamical system is based on trajectories of points. C. Schuldt, I. Laptive, and B. Caputo et al. [5] computes local measurements in human actions in terms of spatiotemporal interest points, where features are extracted and can be adapted to the size of moving patterns. I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld et al. [6] learns human action categories using an unsupervised learning method. The probabilistic models is used to extract noisy feature points that arises from a dynamic background. Bay and A.Ess et al. [7] presents a novel scale- and rotation-invariant detector and descriptor, that relies on integral images for image convolutions. Hessian matrix-based measure performs the function of detector, and the descriptor is a distribution based one.

3. Proposed Method

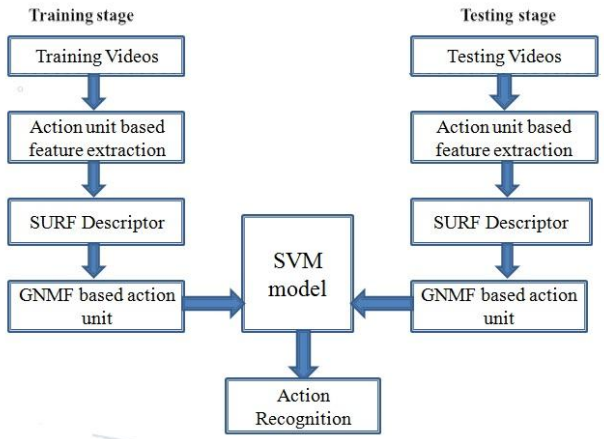

Figure 1: Proposed method for action recognition

Figure 1 shows the proposed model for action recognition. The model mainly consists of two stages. The training stage learns the video for action unit based feature extraction. The testing stage, based on this learned units predicts the action in videos. The SURF descriptor select the discriminative action points for each action class. The GNMF based action unit selects the representative action unit for each action class. SVM is used as the supervised learning algorithm for action prediction.

3.1 Speeded Up Robust Feature Descriptor

We propose a new descriptor called Speeded Up Robust Feature which includes a detector and a descriptor. SURF includes a detector and a descriptor. The detector is a corner detector. The descriptor is a binary step representing the signs of the difference between certain pairs of pixels around the interest point. SURF's detector and descriptor are not only faster, but also more repeatable and distinctive. SURF relies on integral images for image convolution and thus considerably reduces the computation time. Integral images allow for fast and accurate computation of box type convolution filters.

The algorithm has three main steps: interest point detection, local neighborhood description and matching. SURF uses square-shaped filters which forms an approximation of Gaussian smoothing. The sum of the original image within a rectangle can be evaluated quickly using the integral image, requiring evaluations at the rectangle's four corners.

SURF uses a blob detector based on the Hessian matrix to find interestpoints. A measure of local change around the point can be calculated by taking the determinant of the Hessian matrix. Given a point $q = (m, n)$ in an image I, the Hessian matrix $H(q, \sigma)$ at point q and scale σ is given by:

$$H(q,\sigma) = \begin{bmatrix} Lmm(q,\sigma) & Lmn(q,\sigma) \\ Lnm(q,\sigma) & Lnn(q,\sigma) \end{bmatrix} \quad (1)$$

The search for correspondences often requires comparison between images since, interest points can be found at different scales. With the help of a Gaussian filter images are repeatedly smoothed, then they are subsampled to get the next higher level.

3.2 Graph Regularized Non-negative Matrix Factorization

NMF performs learning in Euclidean space. It fails to discover the intrinsic geometrical and discriminating structure of the data. This problem is addressed by using GNMF, which avoids the above limitation by incorporating a geometrically based regularizer. It is not hard to count the arithmetic operations of each iteration in NMF. For GNMF, the matrix used is a sparse matrix. NMF only allows additive combinations between the basis vectors and this property allows NMF to learn a part based representation. The basis vectors learned by GNMF are sparser than those learned by NMF. GNMF have more discriminating power and helps in better representation of actions.

GNMF minimizes the following objective function:

$$S = \left\| Z^i - UV^T \right\| + \gamma Trace(Y^T LY) \quad (2)$$

where U and Y represent two matrices. In case of matrix Y^T each column represents feature points of the corresponding column of Z_i. The specific action units for each action class is obtained by repeating the same procedure. Characteristics of each action class can be learned from the action classes. Each element can be considered as a visual word. A matrix based on the sparsity concept encourages that the samples from the same action classes are constructed using similar action units. The feature points which appear in several intra class samples are suppressed.

The model based on the concept of sparsity helps to find the most discriminative points and can be defined as:

$$min_{S,P^i} \sum_{i=1}^{N} \|Z^i - SP^i\| + \gamma_1 \sum_k \|P_k^i\| + \gamma_2 \|X^i\|_{2,1} \quad (3)$$

where $S = [S_1, S_2, \ldots, S_C]$ represents the dictionary used such that $S_i = [s_{i1}, s_{i2}, \ldots, s_{im_i}]$, where s_{ij} denotes the *j*-th action unit of the *i*-th class.
For a test video *y*, the action unit based feature *x* can be represented by:

$$min\|y - Sx\|_F^2 + \gamma_1\|x\|_{2,1} \quad (4)$$

Where $\|.\|_F$ denotes the Frobenius-norm and $\|.\|_{2,1}$ represents the $l_{2,1}$ norm.

3.3 Support Vector Machine Model

SVM is adopted as the predictive model for discriminative dictionary learning. SVM training algorithm builds a model that assigns new examples to one category or the other. Constructing a hyperplane or a group of hyperplanes in a high-dimensional space, that can be efficiently used for regression and classification can be achieved by using a support vector machine.

By using SVM, a good separation is achieved by the hyperplane that results in the largest distance to the nearest data point of any class. For a SVM classifier larger the margin, lower the generalization error of the classifier. SVM uses the Gaussian kernel with γ^2 distance kernel and is given as:

$$K(H_i, H_j) = exp\left(-\frac{1}{C}\gamma^2(H_i, H_j)\right) \quad (5)$$

where H_i and H_j represents two histograms and *C* is the scale parameter.

4. Simulation Results

Matlab R2015b is used as the platform to perform this task.

.

Figure 2: Input video sequence

The UCF sports dataset is used in the experimental work. The work mainly consists of two stages, testing stage and training stage. Training stage includes, training the machine using a set of videos. Figure 2 shows the input video to be processed for action detection. It is segmented into a large number of frames.

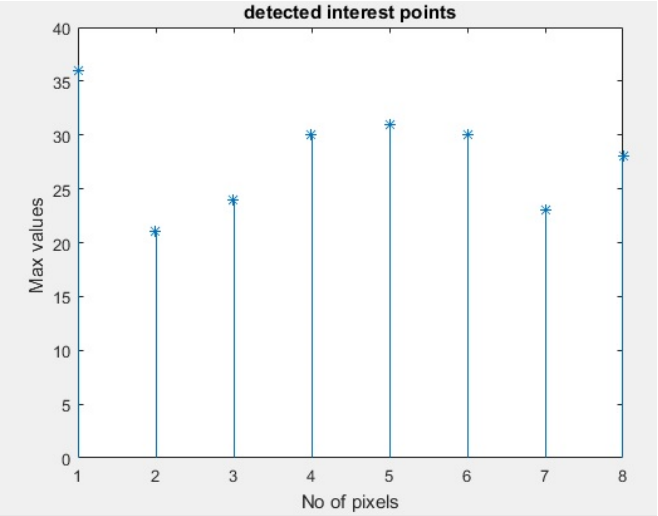

Figure 3: Detected interest points

For each video a corresponding matrix is generated in the work space which shows a large number of intensity point values in the frame. For the set of testing videos the corresponding interest points are detected as shown in Figure 3.

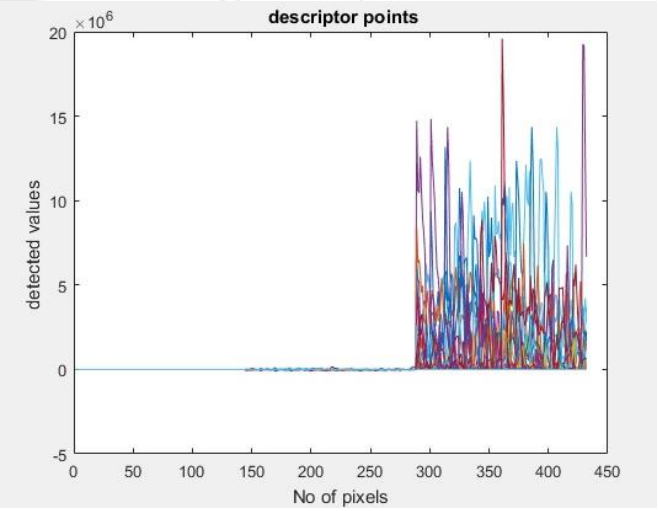

Figure 4: Detected descriptor points

The corresponding descriptor points are also plotted as shown in Figure 4.

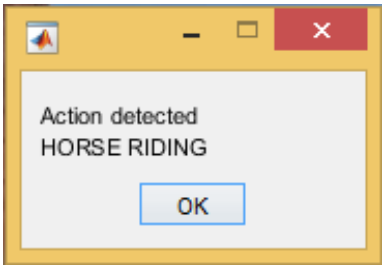

Figure 5: Action Detection

GNMF based matrix factorization and sparsity concept select discriminative action units to detect action. SVM model effectively predicts the action using a supervised learning method. Finally the action is detected as shown in Figure 5.

Table 1: Comparison of time required for action recognition

Descriptor used	*Time(sec)*
LWWC	468
SURF	45

From Table 1 it is clear that SURF descriptor take much less time for action recognition than a LWWC(Locally Weighted Word Context) descriptor when we are using the same concept for action recognition.

5. Conclusion

Action recognition has a wide range of applications. The model based on action units based feature extraction can represent human actions in videos more discriminatively. SURF descriptor improves the discriminability and repeatability of the traditionally used descriptor. The graph regularized nonnegative matrix factorization, avoids local features and learned class specific action units. SVM model effectively predict the actions in videos. Simulation results show that the use of SURF descriptor effectively reduces the computation time for action recognition.

6. Acknowledgment

I would like to express profound gratitude to our Head of the Department, Prof. Sangeeta T. R. , for her encouragement and for providing all facilities for my work. I express my highest regard and sincere thanks to my guide, Asst. Prof. Ms. Amrutha V. Nair, who provided the necessary guidance and serious advice for my work.

References

[1] Haoran Wang and Yuan , "Action Recognition Using Nonnegative Action Component Representation and Sparse Basis Selection," IEEE Trans. Image Proc. vol. 23, no. 2, Feb. 2014.
[2] M. Blank, M. Irani, and R. Basni, “Actions as space-time shapes,” in Proc. 10th IEEE ICCV. Vol. 8, no.12, pp. 1395–1402, Oct. 2005.
[3] D. Cai, X.He, J. Han, and T. S. Haung, "Graph regularized nonnegative matrix factorization for data representation," IEEE Trans. Pattern Anal. Mach. Intel., vol. 33, no. 8, pp.1548c1560, Aug. 2010.
[4] S. Ali, A. Basharat, and M. Shah, "Chaotic invariant for human action recognition," in Proc. IEEE 11th ICCV, vol. 134, no. 6, pp. 1–8, Oct. 2007.
[5] C. Shuldt, I.Laptive, and B. Cuputo, “Recognizing human actions: A loca SVM approach,” in Proc. IEEE 17th ICPR, vol. 3, no. 1, pp.32–36, Aug. 2011.
[6] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in Proc. IEEE, Int. Conf. vol. 79, no. 3, June 2008.
[7] Bay, A. Ess, T. Tuytelaars, and L. Van Gool. “SURF: Speeded Up Robust Features,”Computer vision and image understanding (CVIU). vol. 110, no. 3, pp. 346–359, 2008.

Author Profile

Arya P Muraleedharan received B-Tech degree in Electronics and Communication Engineering from M.G University, Kerala at Sree Buddha college of Engineering in 2015. And now she is pursuing her M-Tech degree in Communication Engineering under the APJ Abdul Kalam Technological University at Sree Buddha college of Engineering.

Amrutha V Nair is working as Assistant Professor in department of Electronics and Communication, Sree Buddha college of Engineering, Elavumthitta, Pathanamthitta.