HUMAN ACTION RECOGNITION USING EXTREME LEARNING MACHINE BASED ON VISUAL VOCABULARIES

M.Ramu¹ and Dr.K.Merriliance MSC.,M.Phil.,Ph.D.,²
Department of Computer Applications, Sarah Tucker College, Thirunelveli-7.

Abstract: This project introduces a novel recognition framework for human actions using hybrid features. The hybrid features consist of spatio-temporal and local static features extracted using motion-selectivity attribute of 3D dual-tree complex wavelet transform (3D DT-CWT) and affine SIFT local image detector, respectively. The proposed model offers two core advantages: (1) the framework is significantly faster than traditional approaches due to volumetric processing of images as a ‘3D box of data’ instead of a frame by frame analysis, (2) rich representation of human actions in terms of reduction in artifacts in view of the promising properties of our recently designed full symmetry complex filter banks with better directionality and shift-invariance properties. No assumptions about scene background, location, objects of interest, or point of view information are made whereas bidirectional two-dimensional PCA (2D-PCA) is employed for dimensionality reduction which offers enhanced capabilities to preserve structure and correlation amongst neighborhood pixels of a video frame. In recent years, research community has witnessed considerable interest in activity recognition due to its imperative applications in different areas such as human-computer interface, gesture recognition, video indexing and browsing, analysis of sports events and video surveillance. Traditional activity recognition approaches have inherent limitations for robust categorization and localization of actions in presence of non-stationary background, varying posture and size of moving objects, and occlusions. Whereas variations in illumination, behavior and mutual interactions of dynamic objects in scenes add further complications to recognition tasks.

INTRODUCTION

Human activity recognition is one of the main challenging topics in computer vision research. Its importance may be explained by many successes for instance in video games, video surveillance, sport analysis, back and neck pain control, sign language understanding, human-robot interaction, etc. Various types of activities indeed appear in everyday life, such as gestures, actions, human-object interactions, social interactions and group activities. Nevertheless, some issues still exist in human action recognition automation: the first ones are inherent to user gesture productions, the other ones come from data acquisitions. The classical recognition methods are often biased by numerous factors: dynamical variations (dynamic and phlegmatic users), temporal variations (slow and fast users), physical variations (device weight, user morphology, left or right-handed users, contextual environment, parallel user activities, etc.), paradigm variations (mono versus multi users, open or closed world paradigm, etc.), and cultural interpretation variations (i.e. one gesture may have different meanings regarding several cultures). Other difficulties impacting the data acquisition system are the presence of occlusions, no rigid motions, view-point changes, background interferences, etc. Therefore, a human action recognition system is designed to deal with these issues. The early methods propose pose estimators as action features and build appearance models based on shape or motion information. Then, in order to avoid segmentation and object tracking issues, local features and bag-of-features representations strategies are proposed. This offers compact action descriptors to be classified as dense trajectories or sparse motion vectors. In contrast, other studies deal with mid-level representations and temporal models. On one side, the main objectives are to group point trajectories into tentative action parts by similarity in motion and appearance and then learn discriminative models with latent assignment of action parts. This allows in particular localizing discriminative action parts.

RELATED WORKS

A detailed study of the existing system is necessary. The functions of the system, requirements for the users, structure of the current system is made through the system study. The problems faced in the current system are found and solution pertaining to it is done in the system study.

In [1], the authors have introduced an action recognition framework that uses concepts from the theory of chaotic systems to model and analyze nonlinear dynamics of human actions. Trajectories of reference joints are used as the representation of the non-linear dynamical system that is generating the action. Each trajectory is then used to reconstruct a phase space of appropriate dimension by employing a delay-embedding scheme. The properties of the reconstructed phase space are captured in terms of dynamical and metric invariants that include Lyapunov exponent, correlation integral and correlation dimension. Finally, the action is represented by a feature vector which is a combination of these invariants over the reconstructed phase space are captured in terms of dynamical and metric invariants that include Lyapunov exponent, correlation integral and correlation dimension. The properties of perfect reconstruction and computational efficiency with good well-balanced frequency responses. No assumptions about scene background, location, objects of interest, or point of view information are made where variations in illumination, behavior and mutual interactions of dynamic objects in scenes add further complications to recognition tasks.

In [2], the authors have proposed human action detection based on a successive convex matching scheme. Human actions are represented as sequences of postures and specific actions are detected in video by matching the time-coupled posture sequences to video frames. The template sequence to video registration is formulated as an optimal matching problem. Instead of directly solving the highly non-convex problem, our method convexifies the matching problem into linear programs and refines the matching result by successively shrinking the trust region. The proposed scheme represents the target point space with small sets of basis points and therefore allows efficient searching. This matching scheme is applied to robustly matching a sequence of coupled binary templates simultaneously in a video sequence with cluttered backgrounds.

In [3], the authors have described a form of discrete wavelet transform, which generates complex coefficients by using a dual tree of wavelet filters to obtain their real and imaginary parts. This introduces limited redundancy (2ⁿ⁻¹ for m-dimensional signals) and allows the transform to provide approximate shift invariance and directionally selective filters (properties lacking in the traditional wavelet transform) while preserving the usual properties of perfect reconstruction and computational efficiency with good well-balanced frequency responses. Here we analyze why the new transform can be designed to be shift invariant and describe how to estimate the accuracy of this approximation and design suitable filters to achieve this. We discuss new variational transforms of the new transform, based on odd/even and quarter-sample shift (Q-shift) filters. We then derive briefly how the dual tree may be extended for images and other multi-dimensional signals, and finally summarize a range of applications of the transform that take advantage of its unique properties.
In [4], the authors have presented a method for extracting distinctive invariant features from images that can be used to perform reliable matching between different views of an object or scene. The features are invariant to image scale and rotation, and are shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination. The features are highly distinctive, in the sense that a single feature can be correctly matched with high probability against a large database of features from many images.

In that paper [5], they have proposed a framework that fuses multiple features for improved action recognition in videos. The fusion of multiple features is important for recognizing actions as often a single feature based representation is not enough to capture the imaging variations (view-point, illumination etc.) and attributes of individuals (size, age, gender etc.). Hence, we use two types of features: i) a quantized vocabulary of local spatio-temporal (ST) volumes (or cuboids), and ii) a quantized vocabulary of spin-images, which aims to capture the shape deformation of the actor by considering actions as 3D objects \((x, y, t)\). To optimally combine these features, we treat different features as nodes in a graph, where weighted edges between the nodes represent the strength of the relationship between entities. The graph is then embedded into a k-dimensional space subject to the criteria that similar nodes have Euclidian coordinates which is closer to each other. This is achieved by converting this constraint into a minimization problem whose solution is the eigenvectors of the graph Laplacian matrix. This procedure is known as Fiedler Embedding.

In [6], they have presented a systematic framework for recognizing realistic actions from videos “in the wild.” Such unconstrained videos are abundant in personal collections as well as on the web. Recognizing action from such videos has not been addressed extensively, primarily due to the tremendous variations that result from camera motion, background clutter, changes in object appearance, and scale, etc. The main challenge is how to extract reliable and informative features from the unconstrained videos. We extract both motion and static features from the videos. Since the raw features of both types are dense yet noisy, we propose strategies to prune these features. We use motion statistics to acquire stable motion features and clean static features. Furthermore, PageRank is used to mine the most informative static features. In order to further construct compact yet discriminative visual vocabularies, a divisive information-theoretic algorithm is employed to group semantically related features. Finally, AdaBoost is chosen to integrate all the heterogeneous yet complementary features for recognition.

In [7], the authors have proposed a novel model for human action categorization. A video sequence is represented as a collection of spatial and spatial-temporal features by extracting static and dynamic interest points. We propose a hierarchical model that can be characterized as a constellation of bag-of-features and that is able to combine both spatial and spatial-temporal features. Given a novel video sequence, the model is able to categorize human actions in a frame-by-frame basis.

In [8], the authors have used a new type of non-separable 3D wavelet transform for video denoising and overcome the motion-mixture problem by using oriented complex wavelets. This wavelet transform is a 3D version of Kingsbury's 1D and 2D dual-tree wavelet transforms. We also investigate video denoising techniques using a combination of both 2D and 3D oriented wavelet transforms. The results are compared with those obtained by separable wavelet transforms.

In [9], the authors have proposed a present a novel unsupervised learning method for human action categories. A video sequence is represented as a collection of spatial-temporal words by extracting space-time interest points. The algorithm automatically learns the probability distributions of the spatial-temporal words and the intermediate topics corresponding to human action categories. This is achieved by using latent topic models such as the probabilistic Latent Semantic Analysis (pLSA) model and Latent Dirichlet Allocation (LDA).

In [10], the authors have presented a novel approach for automatically learning a compact and yet discriminative appearance-based human action model. A video sequence is represented by a bag of spatiotemporal features called video-words by quantizing the extracted 3D interest points (cuboids) from the videos. The proposed approach is able to automatically discover the optimal number of videoword clusters by utilizing Maximization of Mutual Information (MMI). Unlike the k-means algorithm, which is typically used to cluster spatiotemporal cuboids into video words based on their appearance similarity, MMI clustering further groups the video-words, which are highly correlated to some group of actions. To capture the structural information of the learnt optimal video-word clusters, we explore the correlation of the compact video-word clusters. We use the modified correlogram, which is not only translation and rotation invariant, but also somewhat scale invariant.

**METHODOLOGY**

This section describes the methodology of this work and algorithm description. Below diagram explains the system flow diagram.

**MODULES**

- Feature Extraction
- Regression Process
- Classification Stage

**LOCAL BINARY PATTERN**

- Local binary pattern is a very efficient texture operator. Pixels of an image by thresholding the neighborhood of each pixels are result as a binary number. LBP is also improve the recognition accuracy.

**EDGE ORIENTATION HISTOGRAM**

- EOH is used to capture the texture information of an image. EOH is to build a histogram with a directions of the gradients of the edges. The edge detector returns a matrix of the same size of the image with 1 if there is an image and 0 if there is not an image.

**LOCAL PHASE QUANTIZATION**

- Quantization is a compression technique. A range of value to a single quantum value. Each component in the frequency domain by a constant for that components rounded to zero. Video is breaking the picture into discrete block. DCT is used to calculate the frequency component.
REGRESSION STAGE
- K-NN is a lazy learning method for classifying objects. K stands for number of dataset items that are considered for the classification. SVM determines the optimal hyperplane which maximizes the margin. Margin is the distance between the hyperplane and the nearest sample from the hyperplane.

CLASSIFICATION STAGE
- The primary purpose is to work on sequential data. It recognize features independent of timeshift invariance in order to achieve a set of delays are added to the input. The inputs are predicted values from the first stage regression method.

FEATURE REDUCTION
The statistical method widely used to reduce the dimension of feature set is PCA. The new principal component variables are referred to as new domain metrics. Assuming Z(a-b) matrix as the standardized word vector data with "a" reviews and "b" product attributes. The algorithm for carrying out PCA is:

1. Determine the covariance matrix R of Z.
2. Determine Eigenvalues \( k \) and Eigenvectors \( e_k \).
3. The dimensionality of the data is scaled down.
4. Determine domain metrics for each product review using \( t_k \)

The identification of PCA with its respective features (185, 237 and 261 in our case) for each model is identified by Weka tool. ‘Eigen value >1’ is the stopping rule used for this process. As stopping rule has been imposed, the number of principal components for data model.

Principal Component Analysis Algorithm
1. Find the mean vector.
2. Assemble all the data samples in a mean adjusted matrix.
3. Create the covariance matrix.
4. Compute the Eigenvectors and Eigenvalues.
5. Compute the basis vectors.
6. Represent each sample as a linear combination of basis vectors.

KNN- CLASSIFICATION
K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970’s as a non-parametric technique.

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If \( K = 1 \), then the case is simply assigned to the class of its nearest neighbor.

CONCLUSION
A two stage architecture that combines a simple regression and SIFT and PCA approach. In the second stage SIFT-PCA approach that achieves an improvement performance. Predicted labels only depend on the past information.
REFERENCES