Discovering the Most Influential Human Action Using Web Based Classifier

Soumya R, R.Gnanakumari

Abstract- Vision-based human action recognition is the process of tagging image sequences with action labels. The identification of movement can be performed at various levels of abstraction. In existing system, after collecting an preliminary image set for each action by querying the Web, fit a logistic regression classifier to distinguish the foreground features of the correlated action from the background. In the action recognition process, PbHOG features can be used, which are more robust to the background clutter and variance of the domain. Using the initial classifier, incrementally collect more action images and, at the same time improve the model. Use nonnegative matrix factorization on this set to find the diverse pose clusters for that action and train separate local action classifiers for each cluster of poses. In proposed system it can be done by event monitoring to discover the most influential ordered pair for the human specific action. To this end, it make use of already annotated motion capture datasets and prepare action segmentation as a weakly supervised temporal clustering problem for an unknown number of clusters. Use the annotations to learn a distance metric for skeleton motion using relative comparisons in the form of samples of the same action are more similar than they are to a different action. The learned distance metric is then used to cluster the test sequences. To this end, we employ a hierarchical Dirichlet process that also estimates the number of clusters.

Keywords-Action recognition, Heirarchical Dirichlet process

1.INTRODUCTION

Vision-based human action recognition is the process of tagging images with action labels. Robust solution to this problem has applications in domains like visual surveillance, video retrieval and human–computer interaction. The task is difficult due to variations in motion performance. The task of labeling videos containing human motion with action classes is motivated by many applications both offline and online. Automatic annotation [8] of video enables more efficient searching.

Video annotation is the process of adding interactive commentary to the videos. That is adding background information about the video Image annotation is the process by which a computer system automatically assigns metadata in the form of captionining or keywords to a digital image.In machine learning, unsupervised learning [8] refers to the problem of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution .This distinguishes unsupervised learning from supervised learning. Unsupervised learning is closely related to the problem of density estimation in statistics. However unsupervised learning also encompasses many other techniques that seek to summarize and explain key features of the data. Many methods employed in unsupervised learning are based on data mining method used to preprocess data. Unsupervised learning studies how systems can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of patterns.

The queries can be more naturally specified by the user in case of automatic image annotation. But it is not possible in content-based image retrieval. In the case of CBIR, users requires to search the image concepts such as color and texture.The traditional methods of image retrieval such as those used by libraries have relied on manually annotated images, which is expensive and time-consuming, especially given the large and constantly-growing image databases in existence.

Action recognition can be increased by proposing action pose representation from web,but it needs a large amount of training videos.And it is a challenging process,because it needs to find out large labeled data that covers a diverse set of poses.Action recognition in uncontrolled videos is a difficult task, where it is very tough to find the large amount of necessary training videos to model all the variations of the domain. This problem has been addressed in this paper by proposing a generic method for action recognition. The idea is to use images collected from the Web to discover representations of actions and organize this knowledge to routinely annotate actions in videos. For this purpose, first use an incremental image retrieval procedure to collect and clean up the required training set for constructing the human pose classifiers. The approach is unsupervised because it require no human interference other than simply text querying the name of the action to an

Manuscript received June, 2013.

Soumya R *PG Scholar, Computer Science and Engineering, Coimbatore Institute of Engineering and Technology., Narasipuram, Coimbatore, Tamil Nadu, ,India, 9895426268*

R.Gnanakumari*, Assistant Professor, Computer Science and Engineering, Coimbatore Institute of Engineering and Technology, Narasipuram,Coimbatore,TamilNadu,,India.*

internet search engine. Its benefit is two- fold: 1) improve retrieval of action images, 2) collect a large generic database of action poses, which can then be used in categorization of videos. And how the Web-based pose classifiers can be utilized in conjunction with limited labeled videos can be explored. Ordered pose pairs(OPP) can be used for encoding the temporal ordering of poses in action model. Temporal ordering of pose pairs can increase action recognition accuracy. Selecting the key poses with the help of Web-based classifiers, the categorization time can be cheap. Our experiments demonstrate that, with or without avail-able video data, the pose models learned from the Web can improve the performance of the action recognition systems.

First is proposing a system which incrementally collects action images and videos from the Web by simple text querying. Second is building action models by using the noisy set of images in an unsupervised fashion in this present a method for cleaning the results of keyword retrieval, and learn pose models based on this cleaned dataset. Third is proposing PbHOG features, to be used in presence of background clutter that method denoted as an edge detector.Use the probability of boundary (Pb) operator (PbCanny), which is to perform delineating the object boundaries and then used to extract HOG features based on Pb responses. The action models can be used to re-rank retrieved images and improve the retrieval precision.

The action models learned from one set of videos are adapted for recognition in another set of videos using a transfer topic model. Fourth is using the action pose models to annotate human actions in uncontrolled videos (e.g. YouTube videos). The action pose models learnt from the Web can be used for locating the distinctive poses inside the videos, and further, improve the action recognition. This key pose selection scheme also reduces the training time to a great extent. Fifth is using collected image data from the Web jointly with video data for improving action recognition. Sixth is proposing the OPP method for temporal reasoning about body poses within each action; and using Web-based pose classifiers for selecting the key poses from human tracks for efficient training. The proposed OPP descriptor takes one step further and models the temporal relationships between poses. By this, actions that share similar intermediate poses can be more accurately discriminated.

The main contributions are:

- Proposing a system which incrementally collect action images from the web by simply text querying.
- Building action models by using the noisy set of images in an unsupervised fashion, and
- Using the models to annotate human actions in uncontrolled videos, such as YouTube videos.

2. RELATED WORK

Action recognition [3] can be achieved using local dimensions in terms of spatiotemporal interest points. In spatial recognition, local features have recently been joint with SVM in a robust classification approach. In a similar manner, here, investigate the combination of local space-time features and SVM and apply the resulting approach to the recognition of human actions. Typical scenarios include scenes with cluttered, moving backgrounds, nonstationary camera, scale variations, individual variations in appearance and cloth of people, changes in light and view point and so forth. All of these conditions introduce challenging problems that have been addressed in computer vision in the past.

Recognizing human action [2] is a key component in many computer vision applications, such as video surveillance, human-computer interface, video indexing and browsing, recognition of gestures, analysis of sports events and dance choreography. Some of the recent works done in the area of action recognition have shown that it is useful to analyze actions by looking at the video sequence as a space-time intensity volume. Analyzing actions directly in the space-time volume avoids some limitations of traditional approaches that involve key frames.

To automatically categorize or localize different actions [8] in video sequences is very useful for a variety of tasks, such as video surveillance, object-level video summarization, video indexing, digital library organization, etc. However, it remains a challenging task for computers to achieve robust action recognition due to cluttered background, camera motion, occlusion, and geometric and photometric variances of objects. In this paper, present an algorithm that aims to account for both of these scenarios. A lot of previous work has been presented to address these questions. One popular approach is to apply tracked motion trajectories of body parts to action recognition. This is done with much human supervision and the robustness of the algorithm is highly dependent on the tracking system. Ke et al. apply spatio-temporal volumetric feature that efficiently scan video sequences in space and time. Another approach is to use local space-time patches of videos. Laptev et al. present a space-time interest point detector based on the idea of the Harris and F ¨ *o*rstner interest point operators.

3.IMAGE REPRESENTATION

For training classifiers, a large amount of data is needed, such data is collected manually, which is very costly. The data collected from web are more diverse and less biased than the home-made datasets; therefore it may be more sensible for real-world tasks. Collecting useful training images from the web is difficult due to various challenges. For a given query, the ratio of non relevant images in the retrieved dataset is high. And the relevant image set comprises irregular subsets. For building a consistent training set, each of the subsets should be recognized and represented in the last set. Action images means a set of images in which there is at least one person engaged in a particular action. For a given query the number of non relevant images will be high. Sometimes, more than 50% of the images can be irrelevant. The results of keyword retrieval must be cleaned, and then learn pose models based on these cleaned dataset. After collecting the relevant images, the first step is to extract the location of the human, if no humans are detected in the image, then that image is discarded. A human detector can be used for this purpose, which is effective in detecting [10] people.

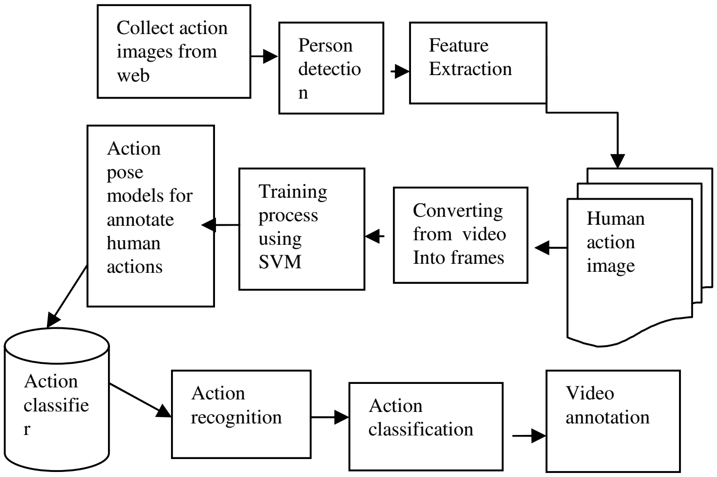
Collect action images from web
Person detectio n
Feature Extraction
Human action image
Converting from video Into frames
Training process using SVM
Action pose models for annotate human actions
Action classifie r
Action recognition
Action classification
Video annotation

Figure1. System Architecture of human action recognition system

Figure 1 shows the system architecture of human action recognition system. First collect action images from web pages by simply text querying the name of the action to a web search engine. From the images extract the person detector, and convert the video into frames. Then the action classifier will classify the actions based on the poses. After identifying and classifying the actions that can be annotated.

3.1 Image collection from webpages

Collecting useful training image datasets from the Web can be difficult due to various challenges. First, for a given keyword- based image search, the ratio of non relevant images in the retrieved dataset tends to be very high. Second, the relevant image set mostly comprises discontinuous subsets, due to different poses, viewpoints and appearances. In order to build a reliable and effective training set, each of these subsets should be identified and represented in the final collected dataset. The number of objects, well as objects' pose and scale vary quite a bit across retrieved images.

3.2 Person Detection

Within the bounding box the detected humans are not always centralized. We can solve this issue via an alignment step based on head area response. Since there is high variance in the limb areas, head detections are the most reliable parts of the detector. The head area should be positioned in the upper center of the bounding box, so for each image we take the detector's output for the head and update the bounding box.

3.3 Feature extraction

once the humans are centralized within the bounding box, extract an image descriptor for each detected area. The descriptor is used to provide a good representation of the poses.For finding the humans from images, Histogram ofOriented Gradients (HOG) is successful. But the clutter in the web images makes it difficult to obtain a pose description. Simple gradient filtering based HOG descriptor is affected by noisy responses.Probability of boundary (pb)operator can be used as an edge detector.

3.4 Testing Input

In this using the training videos and testing the input video using one-vs.-all SVM classifiers over the OPP descriptors. In the SVM classifier,Hollinger kernel method can be used, whose feature map can be explicitly computed by taking the square root of the descriptor values.When video data is available, it is possible to use this video data to improve action models that are learned from Web image data.

3.5 Testing Feature Extraction

Web-based classifiers are effective in selecting the reliable and informative parts of the sequences and use only those detections for action inference.This selection can lessen the testing data size and, hence, reduce the computation time greatly. For this purpose,already trained Web-based pose classifiers can be used.The selected poses and the associated local motion information can further be utilized for efficient action classification.

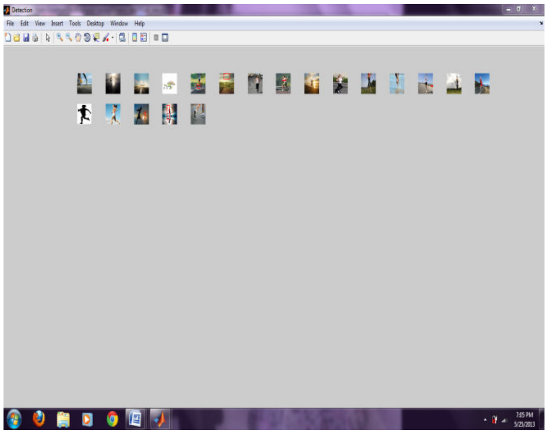

Figure 2.Shows the output of person detector.

3.6 Action classification using classifier NMF

In this using the training videos, learn one-vs.-all SVM classifiers over the OPP descriptors. In the SVM classifier,Hollinger kernel method can be used , whose feature map can be explicitly computed by taking the square root of the descriptor values.when video data is available,use the collected image data together with any available action videos and find out better classifiers over the combined data. Another method is to use, the classifiers learned from Web image data to select the useful parts of the human tracks in videos to facilitate more effective and efficient recognition.

3.7 Metric Learning From poses for Temporal clustering of Human Motions

In this using action labels, constraints can be formulated in terms of similarity and dissimilarity between triplets of feature vectors. Under such constraints, matrix A can be learned by employing Information-Theoretic Metric

Learning (ITML). ITML finds a suitable matrix A by formulating the problem in terms of how similar is A to a given distance parameterized by A0 (typically, the identity or the sample covariance). Provided that coming under equation is a Mahalanobis distance,the problem can be treated as the similarity of two Gaussian distributions parameterized by A and A0 respectively. That leads to an information theoretic objective in terms of the Kullback-Leibler divergence between both Gaussians. This divergence can be expressed as a LogDet divergence, thus yielding the following optimization problem:

$$\begin{gathered} minimize\ D_{ld}(A, A_0) + \lambda D_{ld}\left(diag(\xi), diag(c)\right) \\ s.t.\ \delta_{(i,j)}\left(\xi_{(i,j)} - tr\left(A(X_i - X_j)(X_i - X_j)^T\right)\right) \geq 0 \\ A \geq 0, \xi \geq 0 \qquad (1) \end{gathered}$$

Where D_{ld} is the LogDet divergence, c is the vector of constraints; x is a vector of slack variables (initialized to c and constrained to be component-wise non-negative) that guarantees the existence of a solution and l is a parameter controlling the tradeoff between satisfying the constraints and minimizing the similarity between distances.

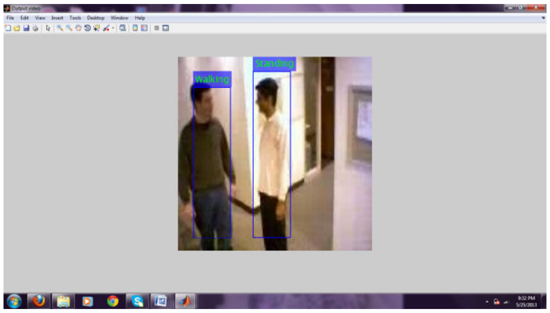

Figure 3.shows annotated video frames

4. PERFORMANCE COMPARISON

To verify the advantages of the proposed work, their performance have to be evaluated. The objective of this section is to compare multiple action with singale action recognition system.

The dataset for the experiment were synthetic dataset. For multiple action, actions like sitting. Jumping and walking were collected. And these were annotated.

Table 1 comparison of Accuracy

Methods	Accuracy (%)
Single Action Recognition	85%
Multiple Action Recognition	93%

5. RESULTS AND DISCUSSIONS

Figure 4 shows the multiple action recognition system based on the parameter accuracy. It is found that accuracy of multiple action recognition is higher than single action recognition system.

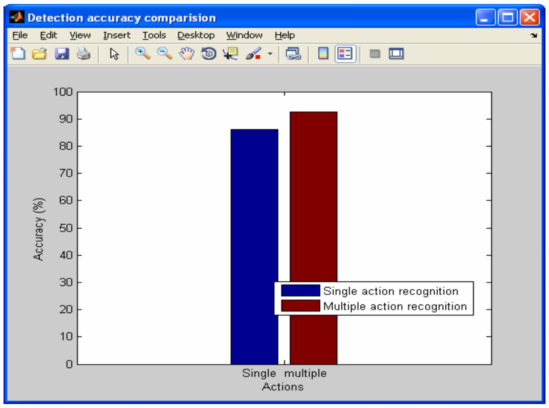

Figure 4. Performance Comparison

6. CONCLUSION

In this paper,the videos are collected from the web and based on the pose the actions are identified and classified.perforformance evaluation shows that multiple action recognition system is having high accuracy when compared to single action system.

7. REFERENCES

1. Adolfo Lopez –Mendez,juragen Gall, Joseph R casas "Metric Learning From Poses For Temporal Clustering Of Human Action".
2. Basri.R , Blank.M, Gorelick.L, Shechtman.E, and Irani.M, (2005) "Actions as space-time shapes," In Proc. ICCV, vol. 2, pp. 1395–140
3. Caputo .B, and Schuldt.C, Laptev.I , (2004) "Recognizing human actions: A local svm approach," in Proc. ICPR, pp. 32–36.
4. Cipolla.R , Kim.T.K, and Wong.S.F,(2007) "Tensor canonical correlation analysis for action classification," presented at the CVPR,Minneapolis,MN.
5. D.Lee and H.Seung, " Algorithms for non-negative matrix factorization",in Proc.NIPS,2001,pp.556-562.
6. D.Tran and A.sorokin,"Human activity recognition with metric learning" in proc ECCV,2008,pp.548-561.
7. F.Schroff,A.Criminisi and A. zisserman,"Harvesting image databases from the web",presented at the ICCV,Rio de Janeiro,Brazil,2007.

8. Fei-Fei.L,Niebles J.C ,and Wang.H,(2006) "Unsupervised learning of human action categories using spatial-temporal words," in Proc.BMVC, pp. 1249–1258.

Soumya R is currently pursuing M.E Computer Science and Engineering at Coimbatore Institute of Engineering and Technology, Coimbatore, Tamil Nadu, (Anna University, Chennai). She completed her B.Tech in Information Technology from M.E.S College of Engineeering ,Kuttipuram,Kerala,(Calicut University, Kerala) in 2010.

R.GnanaKumari is currently Assistant Professor in the Department of Computer Science, at Coimbatore Institute of Engineering and Technology, Coimbatore, Tamil Nadu, (Anna University, Chennai). She completed her B.E in Computer Science and Engineering from Sri Ramakrishna Engineering college, Coimbatore in 2002 and M.E in Computer Science and Engineering from Anna University of Technology in 2011. She has about 3 years experience in industry and 7.6 years experience in teaching.