Survey on Modeling Human object interactions in Still Images

A.N.Bhagat[1], N.B.Pokale[2]

[1,2] Department of Computer Engineering,
TSSM,s Bhivrabai Sawant College Of Engineering and Research,
Narhe, Pune, Maharashtra, India.

Abstract: Human action recognition in still images is attracting much attention in computer vision. This paper considers two action recognition problems in still images. One is the conventional action classification task where a class label is being designed to each action image, and the other is to measure the similarity between action images to model the objects and human poses in images of human actions by using the mutual context model. Recognizing human actions has many applications including video surveillance, human computer inter-faces, sports video analysis and video retrieval. Despite remarkable research efforts and many encouraging advances in the past decade, accurate recognition of the human actions is still a quite challenging task. There are two major issues for human action recognition. One is the sensory input, and the other is the modeling of human actions that are dynamic, ambiguous and interactive with other objects.

Keywords- Human Action Recognition, Action detection

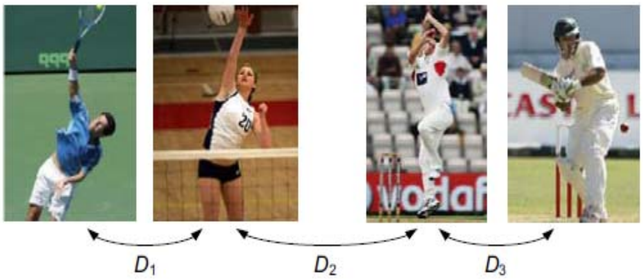
I. INTRODUCTION

Psychologists have proposed that many human-object interaction activities form unique classes of scenes. Recognizing these scenes is important for many social functions. To enable a computer to do this is however a challenging task. If we take people-playing-musical-instrument (PPMI) as an example; to distinguish a person playing violin from a person just holding a violin requires subtle distinction of characteristic image features and feature arrangements that differentiate these two scenes. Most of the existing image representation methods are either too coarse (e.g. Bag of Words) or too sparse (e.g. constellation models) for performing this task.

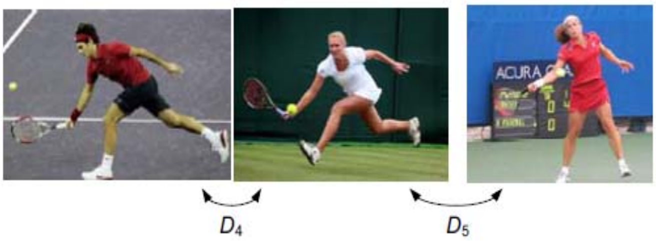
Vision-based human action recognition can be regarded as a combination of feature extraction, and subsequent classification of these image representations. We consider human actions as interactions between humans and objects and jointly model the relationship between them using the mutual context model. This paper shows how the objects and human poses serve as mutual context to facilitate the recognition of each other based on which we address two action recognition tasks:

1. Conventional action classification where we assign a class label to each action image.
2. Measuring the similarity between different action images. The goal is to make the similarity measure consistent with human perception.

Measuring action similarity is very different from conventional action classification problems.



(a) A human action can be more related to some actions than others. $D1 < D2$ because the left-most two images have similar human poses. $D3 < D2$ because the right-most two images are from the same sport and the objects “cricket ball” and “cricket stump” are present in both images.



(b) Human actions lie in a continuous space. Humans are able to capture the difference between different images even if they belong to the same action class. $D4 < D5$ because the left two images have very similar human poses.

II. RELATED WORK

This method builds upon the mutual context model that explores the relationships between objects and human poses in human actions. The model presented in this paper is more flexible and discriminative in that: (1) it learns an overall relationship between different actions, objects, and human poses, rather than modeling each action class separately; (2) it can deal with any number of objects, instead of being limited to the interactions between one human and one object; (3) it incorporates a discriminative action classification component which takes global image information into consideration. While different objects and annotations of action classes can be represented by discrete indexes, human poses lie in a space where the location of body parts changes continuously. To make the joint modeling of actions, objects, and human poses easier, we discretise possible layouts of human body parts into a set of representative poses, termed as atomic poses while poselets are local detectors for specific body parts, the atomic poses consider the whole human body and can be thought of as a dictionary of human poses.

III. Challenges and characteristics of the domain

In human action recognition, the common approach is to extract image features from the video and to issue a corresponding action class label. The classification algorithm is usually learned from training data. The challenges that influence the choice of image representation and classification algorithm.

1. Intra- and inter-class variations

For many actions, there are large variations in performance. For example, walking movements can differ in speed and stride length. Also, there are differences between sizes and proportions of human body individuals. Similar observations can be made for other actions, especially for non-cyclic actions or actions that are adapted to the environment (e.g. avoiding obstacles while walking, or pointing towards a certain location). A good human action recognition approach should be able to generalize over variations within one class and distinguish between actions of different classes. For increasing numbers of action classes, this will be more challenging as the overlap between classes will be higher. In some domains, a distribution over class labels might be a suitable alternative.

2. Environment and recording settings

The environment in which the action performance takes place is an important source of variation in the recording. Person localization might prove harder in cluttered or dynamic environments. Moreover, parts of the person might be occluded in the recording. Lighting conditions can further influence the appearance of the person. The same action, observed from different viewpoints, can lead to very different image observations. Assuming a known camera viewpoint restricts the use to static cameras. When multiple cameras are used, viewpoint problems and issues with occlusion can be alleviated, especially when observations from multiple views can be combined into a consistent representation. Dynamic backgrounds increase the complexity of localizing the person in the image and robustly observing the motion. When using a moving camera, these challenges become even harder. In vision-based human action recognition, all these issues should be addressed explicitly.

3. Temporal variations

Often, actions are assumed to be readily segmented in time. Such an assumption moves the burden of the segmentation from the recognition task, but requires a separate segmentation process to have been employed previously. This might not always be realistic. Also, there can be substantial variation in the rate of performance of an action. The rate at which the action is recorded has an important effect on the temporal extent of an action, especially when motion features are used. A robust human action recognition algorithm should be invariant to different rates of execution.

4. Obtaining and labeling training data

Many works described in this survey use publicly available datasets that are specifically recorded for training and evaluation. This provides a sound mechanism for comparison but the sets often lack some of the earlier mentioned variations. Recently, more realistic datasets have been introduced these contain labeled sequences gathered from movies or web videos. While these sets address common variations, they are still limited in the number of training and test sequences. Also, labeling these sequences is challenging. Several automatic approaches have been proposed, for example using web image search results video subtitles and subtitle to movie script matching present an approach to re-rank automatically extracted and aligned movie samples but manual verification is usually necessary. Also, performance of an action might be perceived differently. A small-scale experiment showed significant disagreement between human labeling and the assumed ground-truth on a common dataset .When no labels are available, an unsupervised approach needs to be pursued but there is no guarantee that the discovered classes are semantically meaningful.

5. Common datasets

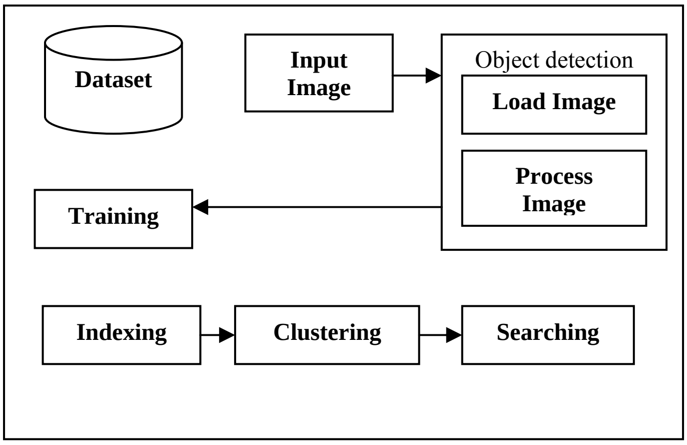
The use of publicly available datasets allows for the comparison of different approaches and gives insight into the (in) abilities of respective methods.

IV. Current System (Methodology)

A set of atomic poses are introduced to learn an overall relationship between different activities, objects, and human poses. The model can deal with the situations where the human interacts with any number of objects (e.g., People interacting with a tennis ball and tennis racket in playing tennis). Also model incorporates a discriminative action classification component and uses state-of-art object and body part detectors which improves the recognition performance.

V. Proposed System

In the proposed work we design the model for detecting an objects and poses of human in the human object interface activities. The following system architecture shows the flow of the proposed work.



Number of input images are loaded and the processing is done to the on all the input image. The process of training, indexing and clustering is done on the input image and saved in the database. Now the new input image loaded and the training, indexing clustering and searching is done. Now we will see in detail each modules of the proposed work.

1. Indexing

In the indexing, training of an image is done. Indexing provides some values to images and according to that values the images gets clustered.

2. Clustering

In the clustering the trained data are placed in the related group without knowing having the advanced knowledge about the group definition. Simple in clustering the partition of a set of data into a group is done. There are number of clustering algorithms are exists like k-means clustering, expectation maximization clustering.

3. Searching

The process of searching is done on the trained image. In the database the training dataset of more than 1000 of images are saved, when the user give the input image for object detection the training of an input image is done after the searching is done. We search the exact match of the input image with the image saved in the database.

VI. Conclusion

The increasing level of sophistication of action recognition algorithms, larger and more complex datasets should direct research efforts to realistic settings. Initially, datasets were not focused on an application domain. However, action recognition in surveillance, human–computer interaction and video retrieval poses different challenges. Human–computer interaction applications require real-time processing, missed detections in surveillance are unacceptable and video retrieval applications often cannot benefit from a controlled setting and require a query interface (e.g. [5]). Currently, there is a shift towards a diversification in datasets. The HOHA dataset [2] targets action recognition in movies, whereas the UFC sport dataset [4] contains sport footage.

Such a diversification is beneficial as it allows for realistic recording settings while focusing on relevant action classes. Moreover, the use of application-specific datasets allows for the use of evaluation metrics that go beyond precision and recall, such as speed of processing or detection accuracy. Still, the compilation or recording of datasets that contain sufficient variation in movements, recording settings and environmental settings remains challenging and should continue to be a topic of discussion.

Related is the issue of labeling data. For increasingly large and complex datasets, manual labeling will become prohibitive. Automatic labeling using video subtitles [1] and movie scripts is possible in some domains, but still requires manual verification. When using an incremental approach to image harvesting such as in the initial set will largely affect the final variety of action performances.

We discussed vision-based human action recognition in this survey but a multi-modal approach could improve recognition in some domains, for example in movie analysis. Also, context such as background, camera motion, interaction between persons and person identity provides informative cues [3].

Given the current state of the art and motivated by the broad range of applications that can benefit from robust human action recognition, it is expected that many of these challenges will be addressed in the near future. This would be a big step towards the fulfillment of the longstanding promise to achieve robust automatic recognition and interpretation of human action.

References

[1]. Sonal Gupta, Raymond J. Mooney, Using closed captions to train activity recognizers that improve video retrieval, in: Proceedings of the Workshop on Visual and Contextual Learning (VCL) at the Conference on Computer Vision and Pattern Recognition (CVPR'09), Miami, FL, June 2009, pp. 1–8.

[2]. Learning realistic human actions from movies, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–8.

[3] Marcin Marszałek, Ivan Laptev, Cordelia Schmid, Actions in context, in:Proceedings of the Conference on Computer Vision and Pattern Recognition(CVPR'09), Miami, FL, June 2009, pp. 1–8.

[4]. Mikel D. Rodriguez, Javed Ahmed, Mubarak Shah, Action MACH: a spatiotemporal maximum average correlation height filter for action recognition, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–8.

[5]. Evan A. Suma, Christopher W. Sinclair, Justin Babbs, Richard Souvenir, A sketch-based approach for detecting common human actions, in:Proceedings of the International Symposium on Advances in VisualComputing (ISVC'08) – part 1, Lecture Notes in Computer Science, LasVegas, NV, December 2008, pp. 418–427 (Number 5358).Ivan Laptev, Marcin Marszałek, Cordelia Schmid, Benjamin Rozenfeld.