ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 3 , March 2015

Moving Object Action Detection Recognition Using HMM Method

Thirumurugan G, Gowri Saranya S

P.G student, Department of CSE,Dhanalakshmi srinivasan Engg College,Perambalur,India.
Assistant Professor, Department of CSE,Dhanalakshmi srinivasan Engg College,Perambalur,India.

ABSTRACT: Recognition of human action is usually addressed in the scope of video interpretation. Meanwhile, common human actions such as "reading a book", "playing a guitar" or "writing notes" also provide a natural description for many still images. Some actions in video such as "taking a photograph" are static by their nature and may require recognition methods based on static cues only. Motivated by the potential impact of recognizing actions in still images and the little attention this problem has received in computer vision so far, To address recognition of human actions in consumer photographs. To construct a new dataset with seven classes of actions in images representing natural variations of human actions in terms of camera view-point, human pose, clothing, occlusions and scene background. The goal of this work is to study recognition of common human actions represented in typical still images such as consumer photographs. To propose a new model for recognizing human attributes (e.g. wearing a suit, sitting, short hair) and actions (e.g. running, riding a horse) in still images. The proposed model relies on a collection of part templates which are learnt discriminatively to explain specific scale-space locations in the images (in human centric coordinates). It avoids the limitations of highly structured models, which consist of a few (i.e. a mixture of) 'average' templates. To learn our model, we propose an algorithm which automatically mines out parts and learns corresponding discriminative templates with their respective locations from a large number of candidate parts.

KEYWORDS: Action detection, HMM method, image processing, Image Acquisition, Human attributes

I. INTRODUCTION

Image processing is a method to convert an image into digital form and perform some operations on it, in order to get an enhanced image or to extract some useful information from it. It is a type of signal dispensation in which input is image, like video frame or photograph and output may be image or characteristics associated with that image. Usually Image Processing system includes treating images as two dimensional signals while applying already set signal processing methods to them. It is among rapidly growing technologies today, with its applications in various aspects of a business. Image Processing forms core research area within engineering and computer science disciplines too. Image processing basically includes the following three steps. First step was importing the image with optical scanner or by digital photography. Second one was analyzing and manipulating the image which includes data compression and image enhancement and spotting patterns that are not to human eyes like satellite photographs. Finally output is the last stage in which result can be altered image or report that is based on image analysis. Image enhancement is the process of adjusting digital images so that the results are more suitable for display or further image analysis. For example, you can remove noise, sharpen, or brighten an image, making it easier to features. Here are some useful examples and methods of image enhancement: Filtering with morphological operators, Histogram equalization and Noise removal.

Camera based text information serves as effective tags or clues for many mobile applications associated with media analysis, content retrieval, scene understanding, and assistant navigation. In natural scene images and videos, text characters and strings usually appear in nearby sign boards and hand-held objects and provide significant knowledge of surrounding environment and objects. Text-based tags are much more applicable than barcode or quick response code. The latter techniques contain limited information and require pre-installed marks. To extract text information by mobile devices from natural scene, automatic and efficient scene text detection and recognition algorithms are essential. However, extracting scene text is a challenging task due to two main factors: First one was cluttered backgrounds with noise and non-text outliers, and finally diverse text patterns such as character types, fonts, and sizes the frequency of occurrence of text in natural scene is very low, and a limited number of text characters are embedded into complex non-text background outliers. Background textures, such as grid, window, and brick, even resemble text characters and strings.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 3 , March 2015

Recognition of human actions is usually addressed in the scope of video interpretation. Meanwhile, common human actions such as “reading a book”, “playing a guitar” or “writing notes” also provide a natural description for many still images. In addition, some actions in video such as “taking a photograph” are static by their nature and may require recognition methods based on static cues only. Motivated by the potential impact of recognizing actions in still images and the little attention this problem has received in computer vision so far, we address recognition of human actions in consumer photographs. To construct a new dataset available with seven classes of actions in 911 Flicker images representing natural variations of human actions in terms of camera view-point, human pose, clothing, occlusions and scene background. To study action recognition in still images using the state-of-the-art bag-of-features methods as well as their combination with the part-based SVM approach. In particular, we investigate the role of background scene context and demonstrate that improved action recognition performance can be achieved by (i) combining the statistical and part-based representations, and (ii) integrating person-centric description with the background scene context. We show results on our newly collected dataset of seven common actions as well as demonstrate improved performance over existing methods on the datasets.

II. RELATED WORK

Describing person attributes is an active research problem in computer vision. Several methods exist in literature, and to tackle the problem of gender recognition. An evaluation of gender classification methods using automatically detected and aligned faces is performed. Interestingly, the evaluation shows that using automatic face alignment methods did not increase the gender classification performance. To propose an approach for gender classification by using facial shape information to construct discriminating models. The facial shapes are represented using 2D fields of facial surface normal. The work of propose to use a part-based approach based on pose lets for describing human attributes. Recently, propose two pose-normalized descriptors based on deformable part models for attribute description. In this paper, to focus on the problem of gender recognition in the wild using semantic information from different body parts. Other than gender recognition, describing actions associated with humans is a difficult problem in computer vision. In action recognition, given the bounding box of a person both at train and test time, the task is to classify the action label associated with each bounding box. Several approaches exist in literature to solve the problem of action recognition.

The bag-of-words based approaches have shown to obtain promising results for action recognition task. To propose an approach based on learning a max margin classifier to learn the discriminative spatial saliency of images. A comprehensive evaluation of colour features and fusion approaches is performed. Besides the bag-of-words framework, several methods have recently been proposed to find human-object interactions for action recognition. A human centric approach is proposed by that works by first localizing a human and then finding an object and its relationship. To introduce an approach based on pose let activation vector that captures the pose in a multiscale fashion. The work of propose a method based on spatial co-occurrences of objects and individual body parts. A discriminative learning procedure is introduced to solve the problem of the large number of possible interaction pairs. Recently, several methods look into combining part-based information within the bag-of-words framework. The work is based on learning a model based on a collection of part templates learnt discriminatively to select scale-space locations in the images. Similarly, this work also investigates how to combine the semantic part-based information within the bag-of-words framework for improved action recognition. In recent years, significant progress has been made in the field of human detection. The part-based approach by shown to provide excellent performance. Besides full-body person detection, localizing specific parts of human body such as face, upper body and hand also exist in literature. A Combining these different body part detectors for human attribute description is an open problem in computer vision. In general, Convolution Neural Networks (CNNs) are only handling the 2D raw inputs. To handle the raw input also for the action recognition, Shuiwang ihave proposed a new CNN model. The motion information that was encoded in multiple adjacent frames was captured by extracting the spatial and temporal features. They have also regularized the outputs with high-level features for boosting up the performance level. The results of proposed work were compared with the publishing methods, which showed the promising results in their proposed work with performance and accuracy for the recognition. A novel view invariant action recognition method was proposed by the authors Anwaar-ul-Haq and they explored the invariance feature of temporal order of action instances. Spatiotemporal features were utilized for ensuring the temporal order during matching. Initially, they had extracted the spatiotemporal features from the video sequences and then fused the features for encapsulating within the class similarity value for the same viewpoints, in which the matching of features across various views were obtained.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 3 , March 2015

III. SYSTEM DESIGN

Automatic image understanding is a major research problem in computer vision. The goal is to analyse an image and be able to make inference based on it e.g. given an image of a person, infer what she is doing. Earlier most of such inference was based on text analysis methods and relied on the (noisy) annotations that came with the image. These annotations could be the tags added by the user or the caption for the image or just the text surrounding the image on a web page. This is changing fast and computer vision technologies that analyse the content of the image, instead of the peripheral noisy text, and make inferences are replacing or augmenting the past systems.

The focus of this project is a semantic description of humans in still images using attributes and actions. Given the daily growing amount of human centric data (e.g. on photo sharing and social networking websites or from surveillance cameras), analysis of humans in images is more important than ever. Most recent work on human attributes or action recognition either rely on, accurate or approximate, estimation of human pose or use general non-human-specific image classification methods. It has been demonstrated that state-of-the-art action recognition can be achieved without solving the difficult problem of pose estimation. Interestingly, several recent methods propose to model interactions between humans and the object(s) associated with the actions. While modelling interactions between humans and contextual objects is an interesting problem, to explore here the broader problem of modelling appearance of humans for attribute and action recognition. Such modelling is critical in the numerous cases where there are no associated objects(e.g. actions like running, walking) and/or the pose is not immediately relevant (e.g. attributes like long hair, wearing tee-shirt).

First, to propose a new image representation to better exploit the class specific spatial information. The standard representation i.e. spatial pyramids, has two shortcomings. It assumes that the distribution of spatial information (i) is uniform and (ii) is same for all tasks. To address these shortcomings by learning the discriminative spatial information for a specific task. To propose a model that adapts the spatial information for each image for a given task. This lends more flexibility to the model and allows for misalignments of discriminative regions e.g. the legs may be at different positions, in different images for running class. Finally, to propose a new descriptor for facial expression analysis. To work in the space of intensity differences of local pixel neighbourhoods and propose to learn the quantization of the space and use higher order statistics of the difference vector to obtain more expressive descriptors.

The capability of a classification system depends on those of its two main components *i.e.* the image representation and the classifier. On one hand, the image representation should lead to similar representations for images of the same class, despite of the intra-class variations, and dissimilar representations for those of different classes, despite of interclass similarity. While on the other, the classifier should be strong enough to perform well, even when the representation scheme is only able to capture the (dis)similaritiesrelativelyweakly.Digital systems, like computers and digital cameras, represent and store images as two-dimensional matrices of pixels where each pixel is a vector of numeric values (usually integers). If the image is gray scale, the vector is one dimensional with the only value indicating the intensity of the pixel (between a fixed minimum values, usually 0, for black and a fixed maximum value, usually 255, for white pixel). If the image has colours, assuming the pixels are coded in Red-Green-Blue (RGB5), the vector is three dimensional with each value similarly indicating the intensity of the red, green and blue colours respectively. The final colour of the pixel is obtained by mixing the RGB colours with those intensities.

A) Image Acquisition

Recognizing actions in still images has recently gained attention in the vision community due to its large applicability to various domains. As opposed to motion and appearance in videos, still images convey the action information via the pose of the person and the surrounding object/scene context. Objects are especially important cues for identifying the type of the action. Previous studies verify this observation and show that identification of objects play an important role in action recognition. A classical way to approach the problem of action recognition in still images is to recover the underlying stick figure. This could be parameterized by the positions of various joints, or equivalently various body parts. In computer graphics this approach has been a resounding success in the form of various techniques for "motion capture". By placing appropriate markers on joints, and using multiple cameras or range sensing devices, the entire kinematic structure of the human body can be detected, localized and tracked over time.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 3 , March 2015

B) Pre-processing

Image pre-processing, also called image restoration, and involves the correction of distortion, degradation, and noise introduced during the imaging process. This process produces a corrected image that is as close as possible, both geometrically and radio metrically, to the radiant energy characteristics of the original scene. Radiometric and geometric are the most common types of errors encountered in remotely sensed imagery. Pre-processing is a common name for operations with images at the lowest level of abstraction for both input and output intensity images. The aim of pre-processing is an improvement of the image data that suppresses unwanted distortions or enhances some image features important for further processing. In this module we convert the RGB image into gray scale images. Then remove the noises from images by using the median filter techniques. The goal of the Median filter is to filter out noise that has corrupted image. It is based on a statistical approach. Typical filters are designed for a desired frequency response. Median filtering is a nonlinear operation often used in image processing to reduce "salt and pepper" noise. A median filter is more effective than convolution when the goal is to simultaneously reduce noise and preserve edges.

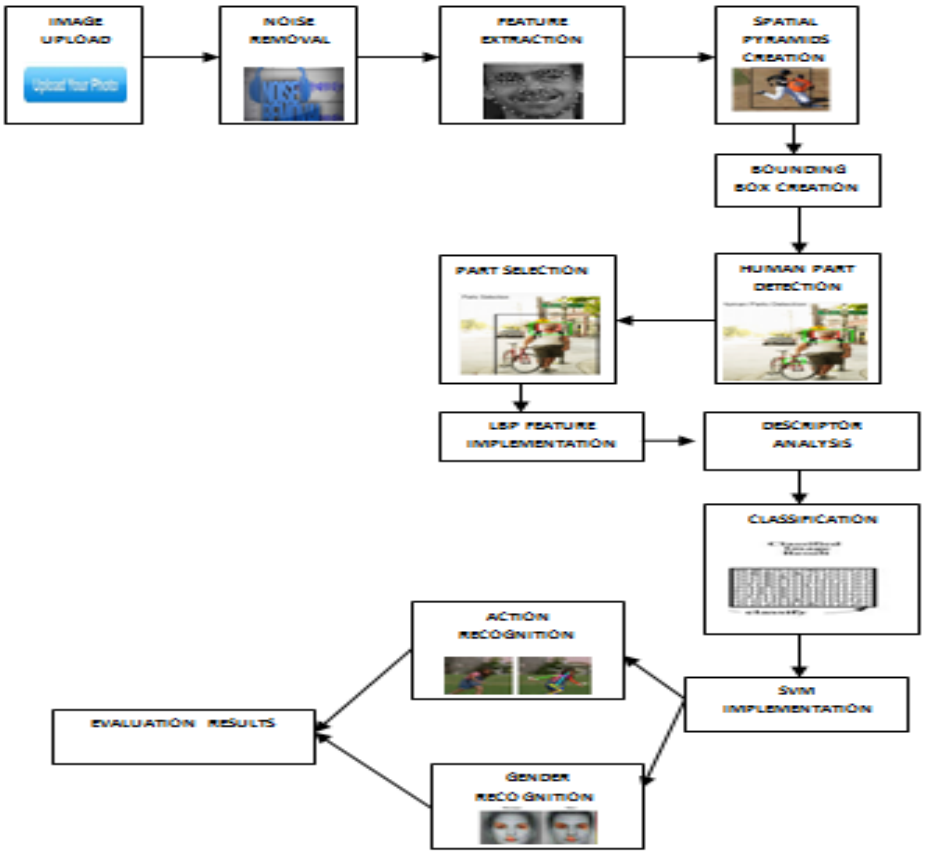

Figure 1: System Architecture Diagram

C) Pyramid Creation

Recognition revolves around relevant features. Here filtering the information from the given data. The right features allow you to compare data in order to determine a difference or a similarity. A feature is a relevant piece of information that helps define or represent a larger whole of data. The feature concept is very general. Some methods use global features, which contain data about the object as a whole. Feature selection can be based on a single frame, but they can also be derived from a sequence of frames. This is done to decouple research in person detection from research in gender/action recognition. The task is then, given a bounding box of the person, to decide on the gender or/and the action of the person. Generally, this is approached by applying a spatial pyramid on the provided bounding box, similar

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 3 , March 2015

as is done for image classification and object detection. The pyramid encodes spatial information and allows the description of features dependent on their relative location in the bounding box. Following this strategy for several features (such as shape, texture, and color) was found. The spatial pyramid allows learning a rough spatial structure of the human outline. But because of the large variety of poses, i.e. people can e.g. be lying, sitting or being captured from the back, the discriminative power of such a representation remains inherently limited. This has recently been acknowledged by research in fine-grained object detection, where the task is to distinguish between hundreds of birds, flowers or airplane models. The exact localization of semantic parts on these classes is considered an important step before going into the feature extraction phase.

D) Part selection

Each of the part detectors fires at multiple locations within the bounding box of the person, yielding a set of hypotheses for all of the parts. These detections come together with a detection score indicating the confidence of the detector. Based on these detectors, we can select human parts. To transform all coordinates to relative coordinates by subtracting the upper left coordinate of the person bounding box, and dividing by the width and respectively height of the bounding box (To indicate relative coordinates). Based on the ground truth bounding boxes we compute the mean location μ_i and its standard deviation σ_i of part i .

E) Classification

A composition of features allows for classification. A composition of features is generated by identifying a stable grouping of features. Classification is the process of coding and organizing the composition of features according to abstract or conceptual descriptions. Compositions are not based on the knowledge of a specific object / action / activity, but rather on general assumptions that hold for most classes. Because classification normally doesn't use specific knowledge, it is a bottom-up process that iterates from low-level grouping to high-level ones. Features are selected and grouped at each level. Grouped or single features or then abstracted to form groups at higher levels. The main problem is what features and how many to group. Grouping to general or to little features may result in classifying different classes as one. Grouping to specific features result in classifying variations of the same class as different classes. The most widely used classifiers are: Neural Network, Support Vector Machines (SVM), k nearest neighbour, Gaussian mixture model, Gaussian, naive Bayes, decision tree and radial basis function classifiers. These classifiers go hand in hand with learning. Implement Local binary patterns (LBP) is the most commonly used feature to extract texture information for image description. The LBP descriptor has shown to obtain state of- the-art results for texture classification. Finally implement Support vector machine classification technique to recognize actions and gender.

IV. CONCLUSION

Human focused visual data makes up a large chunk of the total visual data on the internet and that generated by surveillance. In the near future, it will be critical to have good representations in order to be able to accurately analyse and understand images using automatic computer vision technologies. Analysing human faces would also become an important technology owing to its numerous important applications e.g. surveillance, human computer interaction, medical applications etc. Here presented a new pyramid model for human analysis. The model learns a collection of discriminative templates which can appear at specific scale space positions. It scores a new image by reconstructing it using the available part templates. To propose a stochastic sub-gradient based learning method. The algorithm is capable of exploring a large number of candidate parts and mining out the discriminative parts best suited for the current binary classification. To validated our method on three challenging publicly available datasets for human attributes and actions. To obtained good qualitative and state-of-the art quantitative results, when no external data is used. To analysed the learnt parts with statistics of their discriminative templates and plan to pursue this direction further to gain additional insight. Our future work is human action recognition in video sequences has become an important research topic in computer vision,whose aim is to make machines to recognize human actions using different types of information, especially the motion information, in the video sequences. This research field has captured the attention of several computer science communities due to its strength in providing personalized support for many different applications and its connection to many different fields of study such as medicine, human-computer interaction, or sociology. Three aspects for human activity recognition are addressed including core technology, human activity recognition systems, and applications from low-level to high-level representation. In the core technology, three critical processing stages are thoroughly discussed mainly: human

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 3 , March 2015

object segmentation, feature extraction and representation, activity detection and classification algorithms. In the human activity recognition systems, three main types are mentioned, including single person activity recognition, multiple people interaction and crowd behaviour, and abnormal activity recognition. In future work, to implement this concept in both videos and images under various conditions such as illumination and pose variant conditions and also using improved classifier techniques to classify the human actions accurately.

RFERENCES

[1] L. A. Alexandre, "Gender recognition: A multiscale decision fusion approach," Pattern Recognit. Lett., vol. 31, no. 11, pp. 1422–1427, 2010.
[2] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in Proc. ACM Int. Conf. Image Video Retr.(CIVR), 2007.
[3] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in Proc. Int. Conf. Comput. Vis.(ICCV), 2011.
[4] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in Proc. Int. Conf. Comput. Vis. (ICCV), 2009.
[5] J. Chen et al., "WLD: A robust local image descriptor," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1705–1720,Sep. 2010.
[6] M. Collins, J. Zhang, P. Miller, and H. Wang, "Full body image feature representations for gender profiling," in Proc. Int. Conf. Comput. Vis.(ICCV) Workshop, 2009.
[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. PatternRecognit. (CVPR), Jun. 2005, pp. 886–893.
[8] M. Dantone, J. Gall, G. Fanelli, and L. J. V. Gool, "Real-time facial feature detection using conditional regression forests," in Proc. Comput.Vis. Pattern Recognit.(CVPR), 2012.
[9] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: A study of bag-of-features and part-based representations," in Proc. Brit. Mach. Vis. Conf. (BMVC), 2010.
[10] V. Delaitre, J. Sivic, and I. Laptev, "Learning person-object interactions for action recognition in still images," in Proc. 25th Annu. Conf. NeuralInform. Process.Syst. (NIPS), 2011.
[11] K. Duan, D. Parikh, D. Crandall, and K. Grauman, "Discovering localized attributes for fine-grained recognition," in Proc. Comput.Vis.PatternRecognit. (CVPR), 2012.
[12] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2D articulated human pose estimation and retrieval in (almost) unconstrained still images," Int. J. Comput. Vis., vol. 99, no. 2, pp. 190–214, Feb. 2012.
[13] N. Elfiky, F. S. Khan, J. van de Weijer, and J. Gonzalez, , no. 4, pp. 1627–1636, 2012. "Discriminative compact pyramids for object and scene recognition," Pattern Recognit., vol. 45
[14] R. Farrell, O. Oza, N. Zhang, V. I. Morariu, T. Darrell, and L. S. Davis, "Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance," in Proc. IEEE Int. Conf. Comput. Vis.(ICCV), Nov. 2011, pp. 161–168.
[15] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627–1645, Sep. 2010.