Journal of Quality and Technology Management
Volume X, Issue II, December 2014, Page 155 – 163

A QUALITATIVE ANALYSIS OF FEATURE EXTRACTION BASED ACTION RECOGNITION TECHNIQUES

Y.D. Khan*[1], A. Abid[1], M.S. Farooq[1], K. Abid[2], U. Farooq[1]
[1]University of Management and Technology, Lahore – Pakistan
[2]University of the Punjab, Lahore – Pakistan

ABSTRACT

Action recognition is an interesting problem and has many applications like surveillance, sign recognition, gesture and emotion recognitions. Many solutions to the problem have been suggested by various researchers. In this article a comparative study is performed between such most popular techniques. Various techniques are analyzed which make use of complex models like Discrete Wavelet Transform, Speeded Up Robust Features (SURF), Scale Invariant Feature Transform (SIFT) and image moments. All these models are comparatively analyzed in terms of speed and accuracy. We infer that the moments based algorithm feature extraction method is most balanced in terms of accuracy and efficiency and useful for providing an appropriate quality of results.

INTRODUCTION

Human action recognition is an important field in computer vision. It has wide range of applications such as sign language recognition, keyboard or a remote control, human computer interaction and video retrieval. Now a day's human-machine-interface involves human actions rather than a keyboard, moreover keyboard is being replaced by recognition of sign language. An action is determined by virtue of a sequence of frames rather than a single image as shown in Figure 1. Actions can be represented in form of a combination of body-part movements (Tran, 2012), kinematic features that are derived from some optical flow (Ali, 2010) and sparse representation using spatio-temporaldescriptors (Guha, 2012).

Background noise, camera motion, position and shape of the object are challenging problems. Kinematic features (Ali, 2010) have an issue that is they are not view invariant. An image can be viewed from different

angles which impact differently on an image as well as (Guha, 2012) neglect the spatial and temporal orientation of extracted features so they are unable to detect multiple actions that occur in a video. This problem would be resolved in some extend by using different descriptors or a combination of them. Background noise, camera motion, position and shape of the object are challenging problems. Some researchers present an efficient algorithm for human action recognition by the use of image moments. A collective understanding of image moments describes better information of an image. An image describe a lot of information in sense of its shape, size, position and color in its visual representation but there is another source of information which is quantitative value of a pixel of an image (Ali, 2010), (Cao,2010).

The objective of this research work is to enable the reader to understand the core working of each of these models. Outline of the core of these models is provided later in the article. Furthermore, the results of given analysis in this article eases the choice of the reader in choosing the most appropriate model for an application.

RELATED WORK

Some researchers have used different image moments such as raw moments, centroid moments, and scale invariant moments to yield assiduous results for our problem. These moments make it more affective to understand the position, scale, orientation and size of an object. Bag of Features (BOF) is another approach for classifying of actions (Ullah, 2010). To detect moving objects from complicated backgrounds, Gaussian mixture model (Zhang, 2012) is used, which uses K-means clustering to initialize the model and gets better results for action recognition from videos. Other technique to classify the actions is Color intensity by using motion and shape analysis (Shao, 2012). But there is problem in (Shao, 2012) which occurs while using this approach a region must be selected every time when the scene changes which is an overhead work. HSV-bases segmentation (Busaryev, 2012), silhouette representation. "Envelope Shape" (Zhang, 2011), Motion Energy Image (MEI), and Motion History Image are some techniques for object segmentation. A lot of work is done for action learning or classifying such as multiple instance learning (MIL) (Ali, 2010), BoF approach (Ali, 2010), (Guha, 2012), offline trained classifier (Shao, 2012) and a binary prototype tree (Jiang, 2012).

In this article we present a comparative analysis of various action recognition techniques. Some of the techniques are superior as they recognize an action regardless of its position, scaling, colors and size of objects contained. In some techniques primary purpose is to facilitate video retrieval on the basis of action identification and discovery by providing comprehensive features through video segmentation, feature extraction and feature vector organization. These features are robust to noise because system only tackles these descriptors which are most repeatable and relevant. Computational descriptors become more efficient because only those areas of an image are concerned that contain features are searched. In this article we discuss only those techniques which are enacted on real-time video. The following sections describe these techniques and present a comparison among them.

DISCRETE WAVELET TRANSFORM

Discrete Wavelet Transforms (DWT) are widely used in image applications like face detection, image denoising and image compression. It's a statistical analysis of the signal (image). Discreet Transformation is used in many signal analysis, image processing and image compression. DWT decompose the signals (image) into further sub-bands with time and frequency information and produce high compression ratio. Transformation of signal (image) is another representation of the signal. Figure2 shows this process in detail. DWT analyzes the image with different resolution which gives information of different frequencies within time and space. This technique is basis on the analysis of wavelets which can also be used to analyze other functions. The set of wavelets or a subset is used as a feature vector. The stream of feature vectors is fed into a classifier to search a pattern. An action is flagged by the classifier on successfully identifying a pattern (Khair, 2013), (La-inchua, 2014)

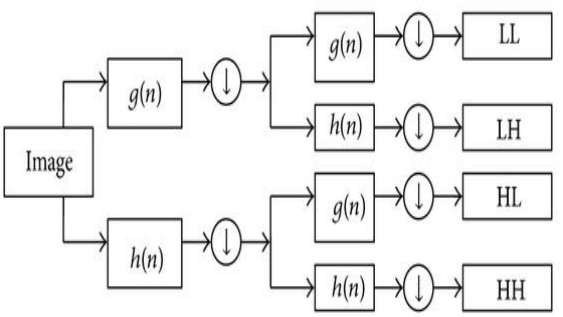

Figure 1: Show a Common 2D Discrete Wavelet Transform

SCALE INVARIANT FEATURE TRANSFORM

Scale Invariant concept was given by David Lowie in 1999. Scale invariant Feature Transformation is used in computer vision algorithms to extract different local features of images to be used in tasks like matching different views of an object or a scene. These features are invariant to the scale, rotation and partially invariant to changes in viewpoints, culture and noise. SIFT key points are extracted from a set of referenced images and stored in database. An object is recognized from the new image by matching its extracted features with the features stored in the database based on the Euclidean Distance. From the full set of matches the key points, which matches the best are filtered out.

This process is divided into following steps:

Scale space:

This step is used to detect those locations and scales that are identifiable from different viewpoints of object. For this scale space function is used:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$$

This step is used to find those points which have low contrast and poor localized at the edges. This is achieved by use in Laplacian. If the function value at z is less than the threshold then this point is removed.(Juan, 2011), (Ali, 2014), (Zhang, 2014).

Orientation Assignment:

This property is used to assign rotation to the key points relative to the local image properties and then achieve invariance to the image rotation. For this compute magnitude and orientation on Gaussian smooth images. Form a histogram from gradient orientation of sample points. Next step is to find highest peak in the histogram and then use this peak with any other peak in the Key point Descriptor.

$$(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2}$$

$$\theta(x, y) = tan^{-1}((L(x, y+1) - L(x, y-1))/(L(x+1, y) - L(x-1, y)))$$

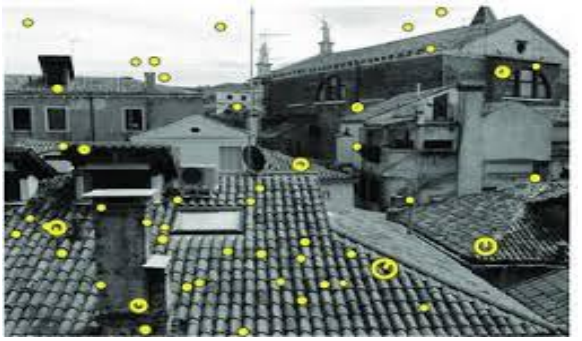

Figure 2 Shows interest points extracted from an image using SIFT

SPEEDED UP ROBUST FEATURES

Speeded Up Robust Features (SURF) is fast interest point detector and descriptor. It is good at handling serious blurring in images and orientation (Rublee, 2011), (Wang, 2013), (Guo, 2014). It consists of following steps.

- Interest point detection
- Descriptors definition with each interest point
- Descriptor Matching

Figure 3: Shows interest points extracted from an image using SURF.

Image moments

Invariant moments are those features which are insensitive to various deformations in the shape and can be used for comparison between two images. So they are used to compute image feature regardless of size, position and rotation. We use three methods to calculate moments. There corresponding equations implements to calculate moments. All the

calculated moments are store in Training Vector. Training Vector is a vector which will save these for moments RNN.

Raw Moments

Raw moments are those moments which are calculated along the origin. If we define an image as function $\boldsymbol{f(x,y)}$where$\boldsymbol{(x,y)}$are the coordinates of the image. For calculating moments a random X variable will be chosen form the image.

The probability distribution function is given as :

$$\boldsymbol{f(X)} = \begin{cases} 1 & if\ f(x,y) \in object \\ 0 & otherwise\ \ background \end{cases}$$

For 2D images continuous function $\boldsymbol{f(x,y)}$ the moment of order $\boldsymbol{(p+q)}$is defined as

$$\boldsymbol{M_{pq}} = \iint_{-\infty}^{\infty} \boldsymbol{x^p\, y^q\, f(x,y) dx dy} \quad (\mathbf{1})$$

Where $\boldsymbol{f(x,y) = x^i\, y^j}$ and **p, q** are the **pth and qth** indices of the moments.The integration is calculated over the area of object. Each other pixel base features other than gray level would be consider as moment. The moment at $\boldsymbol{M_{00}}$will calculate the area of object.

Central Moments

Those moments which are invariant to translation motions are called central moments. From the equation of raw moments central moments can be calculated so the first two order moments from equation (1) $M_{10}M_{01}$are used to locate the center of the mass of the object.

If f(x, y) is an image, then using the center of gravity coordinates $\bar{x}$ and $\bar{y}$the central moments are formally given as

$$\boldsymbol{\mu_{pq}} = \sum_{x} \sum_{y} \boldsymbol{(x-\bar{x})^p (y-\bar{y})^q f(x,y)}$$

The main advantage of central moments is their invariances to translations of the object. Therefore they are suited well to describe the form of the object (Khan, 2014).

Scale Invariant Moments

The raw moments and the centered moments depend on the size of object. This is creates a problem when same object compared but both are captured from different distances. The resolution to this problem is sought by scale invariant moments. Scale invariant moments denoted by μ_{ij} are derived such that the scale or size of the object has no effect. The ratio of corresponding translation invariant central moment with the $(00)^{th}$ central moment provides scale invariant moment as the $(00)^{th}$ central moment signifies the translation invariant area of the image

Rotational Invariant Moments

Rotational moments are those moments which are invariant under translation, changes in scale, and also rotation. Most frequently used are the Hu set of invariant moments.

CONCLUSION AND DISCUSSION

All the above described techniques for action recognition extract the kinematic features of an action. For this purpose a video is split into frames. Features are extracted for each frame in the order of occurrence. This stream of features is fed into a classifier. The classifier based on the pattern of the sequential input classifies the action. In our study we have established that SIFT and DWT are very computational intensive. For the reason of their complexity they cannot be used for live high resolution stream. Moreover SURF is a faster technique. But since it does not yield a fixed interest points therefore it bears problems while using classifiers life neural networks. Image Moments seems to be quite robust technique since it yields a fixed number of features for any image or video. This renders it useful for neural network and equivalent classifiers.

REFERENCES

Ali, S., & Shah, M. (2010). Human action recognition in videos using kinematic features and multiple instance learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *32*(2), 288-303.

Ullah, M. M., Parizi, S. N., & Laptev, I. (2010). Improving bag-of-features action recognition with non-local cues. In *BMVC* (Vol. 10, pp. 95-1).

Busaryev, O., & Doolittle, J. Gesture Recognition with Applications.,2012.

Cao, L., Tian, Y., Liu, Z., Yao, B., Zhang, Z., & Huang, T. S. (2010, July). Action detection using multiple spatial-temporal interest point features. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on* (pp. 340-345). IEEE.

Zhang, F., Wang, Y., & Zhang, Z. (2011, November). View-invariant action recognition in surveillance videos. In *Pattern Recognition (ACPR), 2011 First Asian Conference on* (pp. 580-583).IEEE.

Tran, K. N., Kakadiaris, I. A., & Shah, S. K. (2012). Part-based motion descriptor image for human action recognition. *Pattern Recognition*, *45*(7), 2562-2572.

Guha, T., & Ward, R. K. (2012). Learning sparse representations for human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(8), 1576-1588.

Shao, L., Ji, L., Liu, Y., & Zhang, J. (2012). Human action segmentation and recognition via motion and shape analysis. *Pattern Recognition Letters*, *33*(4), 438-445.

Jiang, Z., Lin, Z., & Davis, L. S. (2012). Recognizing human actions by learning and matching shape-motion prototype trees. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(3), 533-547.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011, November). ORB: an efficient alternative to SIFT or SURF. In *Computer Vision (ICCV), 2011 IEEE International Conference on* (pp. 2564-2571).IEEE.

Juan, L., & Gwun, O. (2009). A comparison of sift, pca-sift and surf. *International Journal of Image Processing (IJIP)*, *3*(4), 143-152.

Khan, Yaser Daanial, et al. "An Efficient Algorithm for Recognition of Human Actions." *The Scientific World Journal* 2014 (2014).

Wang, Heng, and Cordelia Schmid. "Action recognition with improved trajectories". *Computer Vision (ICCV), 2013 IEEE International Conference on* IEEE, 2013.

Guo, Guodong, and Alice Lai. "A survey on still image based human action recognition". *Pattern Recognition* (2014).

Khair, Nurnadia M., et al. "Discrete wavelet transform in recognition human emotional movement through knocking". *Control System, Computing and Engineering (ICCSCE), 2013 IEEE International Conference on* IEEE, 2013.

La-inchua, Jaraspat, Sorawat Chivapreecha, and Suttipong Thajchayapong. "Fuzzy logic-based traffic incident detection system with discrete wavelet transform". *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2014 11th International Conference on* IEEE, 2014.

Ali, Khawlah Hussein, and Tianjiang Wang. "Recognition of Human Action and Identification Based on SIFT and Watermark". *Intelligent Computing Methodologies*. Springer International Publishing, 2014. 298-309.

Zhang, Jia-Tao, Ah-Chung Tsoi, and Sio-Long Lo. "Scale Invariant Feature Transform Flow trajectory approach with applications to human action recognition". *Neural Networks (IJCNN), 2014 International Joint Conference on* IEEE, 2014.