**Paper Review Report:**

Recognising human actions by analysing negative spaces, S. A. Rahman, S.-Y Cho and M. K. H. Leung, IET Computer Vision, Vol. 6, no. 3, pages 197-213, May 2012, Citation: **DOI:**[10.1049/iet-cvi.2011.0185](https://doi.org/10.1049/iet-cvi.2011.0185).

**1. Introduction**

This paper begins by introducing the concept of human action recognition as the process of interpreting human behaviour from image sequences of human actions, noting a few applications as well as common problems. The paper addresses various methods proposed for human motion analysis. Examples include computation of optical flow [1] and eigen-shapes of foreground silhouettes ([2, 3]). However, these methods still face various issues, such as a difficult derivation and a lack of motion information. This paper points out that region-based techniques are much better. Thus, this paper proposes a region-based method that utilises the negative space surrounding a human body in an image to derive action/pose descriptions of human shapes that are less sensitive to foreground segmentation errors.

**2. Proposed Method**

The proposed method is composed of five stages. These include frame grabbing and background segmentation, speed and bounding box computation, region partitioning, pose description/feature extraction and template matching.

**2.1. Frame grabbing and background segmentation**

The proposed method takes video data as an input and therefore to obtain images the video inputs are converted into frames of images. The assumption made by the proposed method is that each video has only one human action being conducted within it (i.e., involving single-person recognition). The next step is background segmentation, which aims to segment an image into a foreground (in this case it would be the human in the image) and a background (everything that is not the human). The background segmentation technique used in this paper is from [4].

**2.2. Speed and bounding box computation**

In order to group the various actions of the person in the image frames, the horizontal and vertical speeds are calculated. To compute the horizontal speed between two frames, the absolute value of the horizontal displacement taken from the top left corner of the bounding box is multiplied by the frame rate. This value is calculated for every consecutive pair of frames, and then summed. The summed value is then divided by the sum of the heights of all the bounding boxes of all frames. To calculate vertical speed the same procedure is used, but with vertical displacement instead. In the next step the smallest upright bounding box which can contain the human person is created to capture the negative space regions. As actions are analysed from right to left, with all frames containing actions moving from left to right are reflected along the vertical axis.

**2.3. Region partitioning**

This stage aims to solve the problem of different numbers of negative space regions being present in multiple poses of the same person conducting the same action captured at different times. To do so regions are partitioned along peninsulas (corresponding to parts of the human body) which point to the bounding box. The process is as follows: for negative space region *n* in a bounding box the position of the closest peninsula to the side of the bounding box that is along the negative space region is determined. Once it is determined the intersection distance between the closest peninsula and the side of the bounding box is found. Then the maximum distances from the side of the bounding box and the human silhouette for both the top half and bottom half of negative space region *n* are determined. The maximum distances are then averaged and divided by the intersection distance to obtain the value of a variable called ‘protrusive’. If ‘protrusive’ is greater than or equal to some minimum threshold then a partition is formed in region *n* at theposition of the closest peninsula.

**2.4. Pose description/ Feature extraction**

Once the negative space regions of the bounding box are partitioned, features need to be extracted in order to compare poses. The two types of features that are extracted include positional and regional-based features. Positional features are extracted by placing 14 anchor points along the sides of the bounding box. Once this is done, every negative space region in the bounding box is associated with one of the anchor points by locating the mid-point that bisects the boundary on the bounding box, and then assigning the anchor point that is closest to the bisecting point of every region. The region-based features that are extracted include the area of the negative space regions (approximated as either triangles or quadrilaterals), eccentricity, orientation, rectangularity (i.e., ratio of region area over the bounding rectangle area of that region). Additionally, they include horizontal and vertical side lengths of the bounding box that is part of the sides of a negative space region.

**2.5. Template matching**

In order to classify the input image sequences the dynamic time warping algorithm is used. It does this by ‘warping’ the sequences non-linearly in the time dimension to determine a measure of their similarity.

**2.6. Novelty of the proposed method**

The novelty of the proposed method is seen through the useof negative space regions to determine the actions of an image sequence. As mentioned in this paper most region-based techniques work with regions that are a part of the human body in the image. Moreover, little work has been conducted on methods utilising negative space regions, with one of the few examples being [5].

**3. Experimental Results**

The datasets used to test the effectiveness of the proposed method include the Weizmann human action (WHA) [6] and KTH [7] datasets. The leave-one-out (LOO) testing scheme is used for both datasets. When testing the classification accuracy of the proposed method, the results showed an accuracy of 100% and 94.67% for the WHA and KTH datasets, respectively. When compared to other methods, the proposed method matched or outperformed all the other methods for WHA. In particular, methods from [8] and [9] which themselves had accuracies over 95% were outperformed. For the KTH dataset, the proposed method outperformed half of the benchmark methods, namely [10] and [11]. The descriptors used are computationally inexpensive. For un-optimized MATLAB code, the time taken was 4.31s on a Pentium 4, 3.4 GHz machine which was much faster than other methods like [12] which took 30s on a similar setup. Further experimentation also showed that the proposed method was robust to noisy segmentation and relatively insensitive to partial occlusion and non-rigid deformation.

**4. Conclusion / Future works**

In conclusion, the proposed method does seem better than other methods not only because it outperformed many of them in terms of classification accuracy, but also because it is computationally efficient and robust. Considering further research, the paper suggests that future work will need to focus on multi-person activity recognition, since the proposed method has only considered single-person recognition.

**References:**

[1] Saleemi, I., Hartung, L., Shah, M.: ‘Scene understanding by statistical modeling of motion patterns’. Proc. Computer Vision and Pattern Recognition, 2010, pp. 2069 – 2076

[2] Diaf, A., Ksantini, R., Boufama, B., Benlamri, R.: ‘A novel human motion recognition method based on Eigenspace’. Proc. Lecture Notes in Computer Science, 2010, pp. 167 – 175

[3] Chen, Y., Wu, Q., He, X.: ‘Human action recognition by radon transform’. Proc. Int. Conf. on Data Mining Workshops, 2008, pp. 862 – 868

[4] Zivkovic, Z., van der Heijden, F.: ‘Efficient adaptive density estimation per image pixel for the task of background subtraction’, Pattern Recognition, Lett., 2006, 27, (7), pp. 773 – 780

[5] Vuuren, M.J.V., Jager, G.: ‘Art and image processing’. Proc. Conf. Pattern Recognition Association of South Africa, South Africa, 2001, pp. 23–28

[6] http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html# Database

[7] Schuldt, C., Laptev, I., Caputo, B.: ‘Recognizing human actions: a local SVM approach’. Proc. Int. Conf. on Pattern Recognition, 2004, pp. 32 – 36

[8] Jhuang, H., Serre, T., Wolf, L., Poggio, T.: ‘A biologically inspired system for action recognition’. Proc. Int. Conf. on Computer Vision, 2007, pp. 1 – 8

[9] Li, W., Zhang, Z., Liu, Z.: ‘Graphical modeling and decoding of human actions’. Proc. IEEE Workshop on Multimedia Signal Processing, 2008, pp. 175 – 180

[10] Schindler, K., van Gool, L.: ‘Action snippets: how many frames does human action recognition require?’. Proc. Computer Vision and Pattern Recognition, 2008, pp. 1 – 8

[11] Ahmad, M., Lee, S.-W.: ‘Human action recognition using shape and CLG-motion flow from multi-view image sequences’, Pattern Recognit., 2008, 41, (7), pp. 2237 – 2252

[12] Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: ‘Actions as space-time shapes’, IEEE Trans. Pattern Anal. Mach. Intell., 2007, 29, (12), pp. 2247–2253