Toward Robust Action Retrieval in Video

Samy Sadek, Ayoub Al-Hamadi, Bernd Michaelis
http://www.iesk.ovgu.de
Institute for Electronics, Signal Processing, and Communications
Otto-von-Guericke University Magdeburg, Magdeburg, Germany

Usama Sayed
http://www.aun.edu.eg
Electrical Engineering Department
Assiut University, Assiut, Egypt

Visual recognition and interpretation of human-induced actions and events are among the most active research areas in computer vision, pattern recognition, and image understanding [4]. In this work, we develop such a framework for robustly recognizing human actions in video sequences. The contribution of the paper is twofold. First a reliable neural model, the Multi-level Sigmoidal Neural Network (MSNN) as a classifier for the task of action recognition is presented. Second we unfold how the temporal shape variations can be accurately captured based on both temporal self-similarities and fuzzy log-polar histograms.

Neural network classifier has many advantages over other competitive machine learning classifiers. Some of these advantages include the high rapidity, easiness of training, realistic generalization capability, high selectivity and great capability to create arbitrary partitions of feature space. However, the neural model, in the standard form, might have low classification accuracy and poor generalization properties because its neurons employ a standard bi-level function that gives only two values (i.e., binary responses) [2]. To relax this restriction and allow the neurons to generate multiple responses, a new functional extension for the standard sigmoidal functions should be developed. This extension is termed the Multi-level Activation Function, and the model that uses this extension is termed the Multi-level Sigmoidal Neural Network (MSNN). It is straightforward to derive a multi-level version from a given bi-level standard sigmoidal activation function $f(x)$ as follows

$$\phi_r(x) \leftarrow f(x) + (\lambda - 1) f(c) \qquad (1)$$

where λ is an integer index running from 1 to $r-1$; r is the total number of levels, and c is an arbitrary constant.

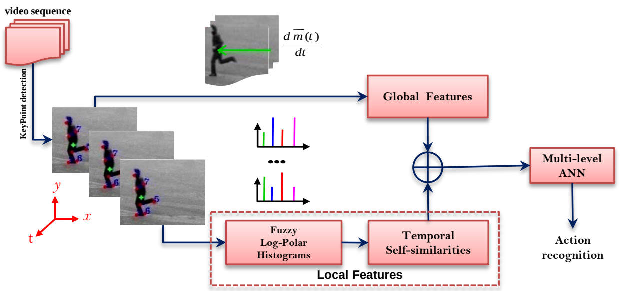

Figure 1: Main components of the action recognizer.

The proposed action recognizer is schematically illustrated in Fig. 1. As seen from the block diagram, the process of recognition systematically works as follows. First keypoints are detected for each video clip (i.e., action snippet), using a Harris based algorithm. To make the method more robust to time warping effects, each action snippet is divided into a number of overlapping time-slices defined by linguistic intervals. Gaussian membership functions are used to describe such intervals:

$$\mu_j(t;\varepsilon_j,\sigma,m) = e^{-\frac{1}{2}\left|\frac{t-\varepsilon_j}{\sigma}\right|^m}, \quad j = 1,2,\ldots,s \qquad (2)$$

where ε_j, σ, and m are the center, width, and fuzzification factor, respectively and s is the total number of time-slices. To extract the local features of the shape representing the action pose at a time, the basic idea of the shape context proposed first by Belongie *et al* [1] should be improved. The idea behind a modified shape context is based on computing rich descriptors for fewer keypoints. Moreover the shape context is reasonably extended by combining local descriptors with the fuzzy memberships functions and temporal self-similarities paradigms. Let B be a set of sampled keypoints $\{(x_i,y_i)\}_{i=1}^n$ representing an action at a time t. Thus log-polar coordinates (ρ_i,η_i) for each keypoint $p_i \in B$, are given by

$$\rho_i = \sqrt{(x_i-x_c)^2+(y_i-y_c)^2},\ \eta_i = \log\left[\arctan\left(\frac{y_i-y_c}{x_i-x_c}\right)\right] \qquad (3)$$

where $\varsigma = (x_c, y_c)$ is the center of mass of B, which is invariant to scaling, rotation or translation. Hence the angle is computed with respect to a horizontal line passing through the center of mass. To calculate the modified shape context of the action pose, the log-polar histogram is overlaid on the shape of action pose. Thus the fuzzy log-polar histogram representing the shape of action at a temporal state j can be constructed by using the membership functions $\mu_j(t)$ as follows

$$h_j(k_1,k_2) = \sum_{\substack{\rho_i \in bin(k_1),\\ \eta_i \in bin(k_2)}} \mu_j(t_i),\ j = 1,2,\ldots,s \qquad (4)$$

where each of these histograms is a 2-d matrix of $d\rho \times d\eta$ dimensional, where $d\rho$ and $d\eta$ are the radial and angular dimensions, respectively. By applying a simple linear transformation on the indices k_1 and k_2, the 2-d histograms is converted into one dimensional (1-d) vectors.

Next the resulting histograms are normalized to the integral value of unity to achieve robustness to scale variations. The normalized histograms obtained can now be used as shape contextual information for classification and matching. Many conventional approaches in various computer vision applications directly combine such normalized histograms to obtain one feature vector per video, which in turn can be classified by any classification algorithm such as Bayes decision rule, ANN, SVM, HMM, etc. In contrast, in this research, we aim to enrich these histograms with the self-similarity analysis. Given a histogram series $H = \{h_1,h_2,\ldots,h_m\}$ that temporally represents m temporal segments of an action, then the temporal self-similarity matrix is given as

$$S = [s_{ij}]_{i,j=1}^m = \begin{pmatrix} 0 & s_{12} & \cdots & s_{1m} \\ s_{21} & 0 & \cdots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \cdots & 0 \end{pmatrix} \qquad (5)$$

where $s_{ij} = s(h_i,h_j)$ is the similarity between histograms i and j. The diagonal entries are zero, because $s(h_i,h_i) = 0\ \forall i$. Meanwhile, because $s_{ij} = s_{ji}$, S is a symmetric matrix. Since the global features tend to be relevant to the current recognition task, the final features fed into the MSNN classifier are constructed by combining local and global features.

During the learning stage, the MSNN classifier is trained using the features extracted from the action snippets in the training dataset. The up diagonal elements of the similarity matrix representing the local features are first transformed into plain vectors, and then fused with the global features. All feature vectors are finally fed into the MSNN classifier to distinguish all action classes. After the learning stage is finished, the system is able to recognize and identify unseen actions.

The overall conclusion is that the incorporation of the fuzzy membership functions makes the method quite robust to shape deformations and time wrapping effects. When validated on the papular KTH dataset [3], the results obtained show the superiority of the method over other popular and recent methods in the literature. Furthermore, the method runs in runtime, and thus can provide timing guarantees to real-time applications.

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on PAMI*, 24 (4):509–524, 2002.

[2] S. Bhattacharyya, P. Dutta, and U. Maulik. Object extraction using self organizing neural network with a multi-level sigmoidal transfer function. In *Proc. 5th International Conference on Advances in Pattern Recognition*, pages 435–438, 2003.

[3] I. Laptev and P. Perez. Retrieving actions in movies. In *Proc. of ICCV*, pages 1–8, 2007.

[4] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2):90–126, 2006.