Action Chart: A Representation for Efficient Recognition of Complex Activities

Hyung Jin Chang
hj.chang@imperial.ac.uk

Department of
Electrical and Electronic Engineering,
Imperial College London, London, UK

Jiyun Kim, Jungchan Cho, Songhwai Oh,
Kwang Moo Yi, Jin Young Choi
{jiyun07,cjc83,songhwai,kmyi,jychoi}@snu.ac.kr

Department of Electrical
and Computer Engineering, ASRI
Seoul National University, Seoul, Korea

Action recognition has been widely studied for decades and there are many successful approaches to recognize relatively simple actions [2]. Recently, more realistic and complex activity recognition tasks have been dealt with, but the current status of the research on complex activities is in its initial phase and far from the recognition ability of human. In our work, we are interested in recognizing temporally very long, complex and diverse action streams. Reliable and efficient recognition methods of this kind of complex action streams can be useful for various tasks, such as complex activity categorization, long video abstraction and similar activity-based video retrieval in YouTube.

In this paper, we propose a pipelined motion-information embedding structure from a high dimensional local feature flow to a low dimensional attentional motion spot flow in order to recognize long and complex action streams efficiently. Figure shows an overall scheme of the proposed method. Each step of the proposed method is focused on extracting distinctive motion information and filtering out noise. In order to reduce high dimensional action video sequences (>640x480x6000 frames) into simple representations while retaining the necessary information, we propose a new composite motion feature generation method by combining various conventional low-level local features. The composite features characterize local, holistic, and sequential motion changes with small memories. The 21-dimensional composite feature sequences are embedded into one-dimensional feature sequences with preserving motion characteristics by proposing a hierarchical embedding method. The one-dimensional embedded feature sequence is utilized to catch distinctive motions. The distinctive motion instances are referred to as attentional motion spots (AMSs), which are automatically determined in our scheme. The AMSs appear in similar feature space-time locations for the same activity classes. We model the distribution of AMSs as a weighted Gaussian mixture model (GMM) using expectation maximization (EM) in embedded feature space-temporal domain. The sketch of this model looks like a music chart, thus we name our representation as *Action Chart*, which is used for action class recognition.

Composite Motion Feature Flow: The composite motion features are newly defined in this subsection by manipulating the low-level local feature information. The composite motion features extracted in each frame form a temporally sequential flow through the whole video frames, referred to as composite motion feature flow (CMFF). The CMFF ($\mathcal{M} = \{\mathcal{M}(t)|t = 1,...,N\}$) is composed of holistic ($\mathcal{M}_H$) and local ($\mathcal{M}_L$) motion features. The holistic motion feature $\mathcal{M}_H$ is composed of five measurements; motion intensity (m_I), motion extent (m_E), motion speed (m_S), motion change (m_C) and motion diversity (m_D). The motion intensity, extent, and speed represent quantitative motion property, and the change and diversity reflect qualitative motion property. At each t frame, the five measurements are obtained independently. The local motion feature ($\mathcal{M}_L(t) \in R^{16}$) represents the relative location distributions of local feature points, $p_{(x,y)}$, using a concentric 16-bin histogram method whose center is placed in the center of human body. We compose the CMFF by concatenating the holistic and local features; $\mathcal{M}(t) = [\mathcal{M}_H(t), \mathcal{M}_L(t)] = [m_I(t), m_E(t), m_S(t), m_C(t), m_D(t), m_L^{[1,...,16]}(t)]^T$ ($\mathcal{M}(t) \in R^{21}$).

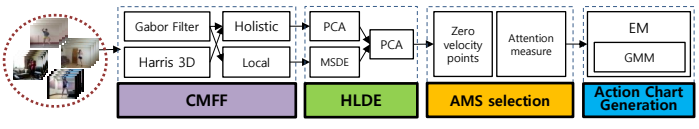

Figure 1: Overall scheme of building *Action Chart*. The proposed method is composed of four steps: (1) composite motion feature flow generation using low-level local features, (2) hierarchical embedding of the feature flow into 1-dimensional (1-D) feature sequence, (3) attentional motion spot selection in the 1-D sequence, and (4) activity modeling and recognition using the attentional motion spots.

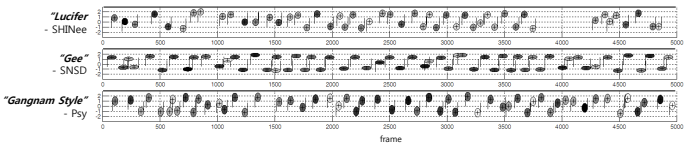

Figure 2: Generated *Action Charts* for the Pop-Dance dataset.

Hierarchical Low Dimensional Embedding: To avoid the curse of dimensionality in analyzing motion streams, it is necessary for the CMFF $\mathcal{M}$ to be embedded to a lower dimensional space. However $\mathcal{M}$ consists of two different information groups; $\mathcal{M}_H$ and $\mathcal{M}_L$. Each dimension of $\mathcal{M}_H$ implies independent and distinctive motion information, while the 16 dimensions of $\mathcal{M}_L$ represent only one motion information as a combination. In other words, the importance of each dimension is different. To handle this problem, we propose a hierarchical low dimensional embedding (HLDE) method.

AMS Selection: By mimicking the human perception mechanism, we propose a method to catch and focus on distinctive instances along the motion feature flow X. We define the distinctive instances as AMS, and we use velocity (the first derivative) of X to find the AMS, which is similar to the human mechanism of using motion changes as a clue for segmentation. The number of zero-velocity points is determined automatically. To avoid the false detection problem that other zero-velocity based methods [3] suffer from, we introduce an attention measure η at j^{th} zero-velocity point z_j, and the η is used for filtering out noisy zero-velocity points by thresholding. The number of attentional points determined automatically.

Action Chart Generation and Recognition: Each AMS set $\mathbf{Y}^c$ follows a weighted GMM distribution in the feature space-time domain. It can then be written as $p(\mathbf{Y}^c|\theta^c) = \sum_{m=1}^{k^c} \omega_m^c p(\mathbf{Y}^c|\theta_m^c)$, where $\omega_1^c, ..., \omega_{k^c}^c$ ($\omega_m^c \geq 0, m = 1, ..., k^c$, and $\sum_{m=1}^{k^c} \omega_m = 1$) are the *weights of each component*, each θ_m^c is a set of Gaussian parameters $\theta_m^c = \{\mu_m^c, \Sigma_m^c\}$ defining m^{th} component, and $\Theta^c \equiv \{\theta_1^c, ..., \theta_{k^c}^c, \omega_1^c, ..., \omega_{k^c}^c\}$. We define the Θ^c as *Action Chart* of action class c. To estimate each Θ^c, the EM algorithm is used. We adopt the Figueiredo and Jain [1] algorithm for unsupervised parameter estimation. The class of the test action stream X^{test} is determined through maximum log-likelihood. The AMS set of X^{test} is obtained and represented as $Y^{test} = \{y_1^{test}, ..., y_{n^{test}}^{test}\}$. The class recognition is performed by matching the Y^{test} and the trained *Action Chart* Θ^c one by one.

In order to validate the proposed method, we generated a new complex action dataset; the Pop-Dance dataset. The experimental results showed that the *Action Chart* could give a promising recognition performance with a very low computational load. Furthermore it could be used for abstracting a long video sequence aims. Our method can contribute to recognizing repetitive sequential activities (e.g. workplace safety, retail fraud detection or sweethearting, and product quality assurance) and sequentially combined action tasks (e.g. sign language and cooking menu), which are our future research.

[1] Mario A.T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):381–396, 2002.

[2] J.K.Aggarwal and M.S.Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3), April 2011.

[3] Daniel Weinland. *Action Representation and Recognition*. PhD thesis, Institut National Polytechnique De Grenoble, oct 2008.