ICA-BASED ACTION RECOGNITION FOR HUMAN-COMPUTER INTERACTION IN DISTURBED BACKGROUNDS

Wei-Yao Chiu and Du-Ming Tsai
Department of Industrial Engineering and Management, Yuan-Ze University, Taiwan, R.O.C.

Keywords: Human-computer Interaction, Action Recognition, Video Surveillance, Spatiotemporal Representation, Independent Component Analysis.

Abstract: In this paper, we propose an independent component analysis (ICA) based scheme for action recognition. The proposed method does not require the feature design or modeling process, and it is computationally very fast for real-time applications. Experimental results show that it is robust on disturbed backgrounds where both foreground and background objects are moving simultaneously in the scene. The experiments also reveal that the proposed method is very effective for human-computer interaction.

1 INTRODUCTION

In this paper, we propose a scheme based on independent component analysis (ICA) for action recognition. It uses the exponential motion history image (EMHI) for spatiotemporal representation of an action, and the discriminant features are then automatically extracted from the EMHI by ICA basis image reconstruction. A complex action can be constructed with only a few training samples of a reference action. ICA is used to generate the basis of action templates from the training EMHI images. Each action to be recognized can be constructed by a linear combination of the ICA basis templates. If the input scene video contains an action similar to the training sample, then the corresponding EMHI can be well reconstructed from the linear combination of the basis templates. The coefficients of the linear combination are used as the discriminant feature vector for action classification. It needs only a few action samples for the training and is robust on disturbed backgrounds with both moving foreground and background objects present simultaneously in the scene. The proposed method can be applied to a moving background environment, such as a street scene with moving vehicles.

2 ICA-BASED APPROACH FOR ACTION RECOGNITION

This section discusses the ICA-based approach for action recognition that comprises the Exponential Motion History Image (EMHI) for spatiotemporal representation of motions, extraction of motion features by ICA, and the distance measure for classification.

2.1 Spatiotemporal Representation

The proposed activity recognition scheme first constructs a global representation of motion that can describe the changes in both temporal and spatial dimensions. In this study, we focus on the recognition of specific activities that can be defined beforehand. We also use the exponential motion history image (EMHI) as the spatiotemporal representation of a video sequence and then apply the ICA basis images for classifying the pre-determined actions. For completeness, the construction of EMHI from a stream of video sequences is briefly described here, and the advantage of EMHI over the conventional MHI for the recognition of specific activities is also analysed.

The existing spatiotemporal representations of motions generally describe the temporal context with a fixed duration in a video sequence. The motion representation from a fixed number of image

Chiu W. and Tsai D..
ICA-BASED ACTION RECOGNITION FOR HUMAN-COMPUTER INTERACTION IN DISTURBED BACKGROUNDS.
DOI: 10.5220/0003888605190522
In *Proceedings of the International Conference on Computer Graphics Theory and Applications* (GRAPP-2012), pages 519-522
ISBN: 978-989-8565-02-0

frames may not sufficiently capture the salient and discriminative properties for a large variety of activities. A short observation duration cannot describe the full cycle of an activity. In contrast, an excessively long observation duration may mix two or more different activities or reduce the significance of a unique activity in the spatiotemporal representation.

Let $f_t(x, y)$ be the *t*-th time frame image in a video sequence. The exponential history image (EMHI) up to time frame *t* is defined as

$$E_t(x, y) = M_t(x, y) + E_{t-1}(x, y) \cdot \gamma \tag{1}$$

where γ is the energy update rate, $0 < \gamma < 1$, and

$$M_t(x, y) = \begin{cases} 1, & \text{if } f_t(x, y) \in \textit{foreground} \\ 0, & \text{otherwise} \end{cases}$$

Since the activity recognition scheme proposed in this study does not rely on the parts of the extracted human body, the accuracy of the segmented foreground region has no significant impact on the EMHI spatiotemporal representation. The initial value of energy is set to zero at time frame 0, i.e., $E_0(x, y) = 0$, for all pixels. The energy of a pixel $E_t(x, y)$ will be increased if it remains as a foreground point up to time *t*. It is decayed only when it becomes a background point. In eq. (1), $M_t(x, y)$ is set to 1 if pixel(x, y) at time frame *t* is a detected foreground point. Otherwise, it is set to 0 for a background point.

To compare the effectiveness of spatiotemporal representations from the proposed Exponential MHI (EMHI) and the conventional MHI, Figure 1 displays the resulting energy images for a set of various actions. Figure 1(a) shows the first frame of each action video sequence in the Weizmann dataset (Gorelick *et al.* 2007). Figure 1(b) shows the corresponding Motion History Image (MHI) of each action, in which the observed duration is set at 50 frames. Figure 1(c) displays the corresponding Exponential Motion History Image (EMHI) of each action, in which the update rate γ is given by 0.99.

The conventional MHI representations are too similar to distinguish the different actions of Jump, Run, Walk and Side-walk. EMHI, however, can generate distinguishable representations for the different repetitive actions. As seen in Figures 1(a2) and (a3) for the Run and Walk actions, the EMHI representations show two actions with different leg-motion patterns. Therefore, EMHI gives a discriminative spatiotemporal representation for describing various action patterns, especially for those that involve prolonged, repetitive motions.

Figures 2(a1)~(a3) show the one-hand wave action on a disturbed background. The objects causing disturbance include many pedestrians, cars, and motorcycles passing through the background. Figures 2(b1)~(b3) present the conventional MHI image with great interference from the background objects. Figures 2(c1)~(c3) illustrate the proposed EMHI image with low energy noise from the moving background objects, wherein the waving motion is still distinctly intensified in the energy image. From the demonstrative samples in Figures 1 and 2, it is clear that the proposed EMHI can represent different activities that involve simple or complex motions on a disturbed background.

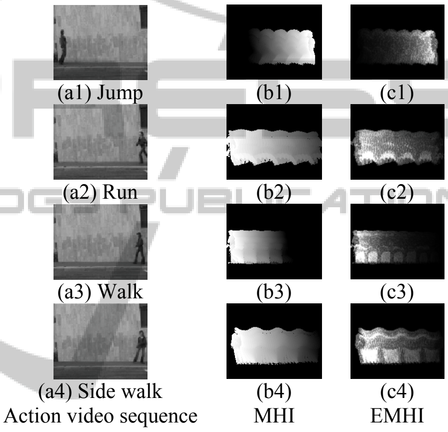

Figure 1: Different activities in the Weizmann dataset and corresponding representations: (a1)-(a4) single person activity; (b1)-(b4) MHI representation; (c1)-(c4) EMHI representation.

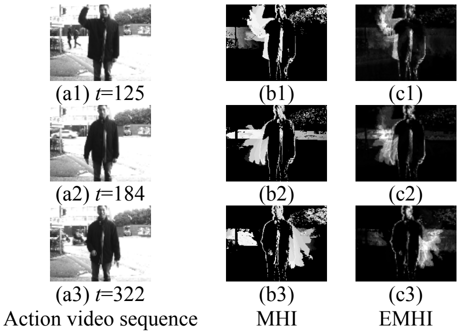

Figure 2: Repetitive one-hand waving on a disturbed background and corresponding representation: (a1)-(a3) single person activity; (b1)-(b3) MHI representation; (c1)-(c3) EMHI representation. (The symbol *t* represents the frame number in the video sequence with fps=15.)

2.2 ICA-based Feature Extraction

The proposed EMHI gives the global spatiotemporal representation of an activity. To construct a classification system for action recognition, the classical approaches need first to design and extract discriminant features from the spatiotemporal representation and then to perform feature selection to find the best feature combination based on specific selection criterion, and finally to apply a classifier to identify individual actions.

In the basic ICA model (Hyvarinen et al., 2000), the observed mixture signals $\mathbf{X}$ can be expressed as

$$\mathbf{X} = \mathbf{AS}$$

where $\mathbf{A}$ is an unknown mixing matrix, and $\mathbf{S}$ represents the latent source signals, meaning that they cannot be directly observed from the mixture signals $\mathbf{X}$. The source signals are assumed to be mutually statistically independent. Based on this assumption, the ICA solution is obtained in an unsupervised learning process that finds a de-mixing matrix $\mathbf{W}$. The matrix $\mathbf{W}$ is used to transform the observed mixture signals $\mathbf{X}$ to yield the independent signals, i.e. $\mathbf{Y} = \mathbf{WX}$. The components of $\mathbf{Y}$, called independent components, must be as mutually independent as possible.

The objective of the algorithm for an ICA model is to maximize the statistical independence (non-Gaussianity) of the ICs. The non-Gaussianity of the ICs can be measured by the negentropy (Hyvarinen and Oja, 1997).

For action recognition, a set of training samples is collected from all reference actions, where each sample is represented by the EMHI. Let the training samples involve *K* specific actions and each reference action include more than one EMHI sample. To find the basis images of multiple reference actions using ICA, let $\mathbf{X} = \left\{ \mathbf{x}_i^j, i = 1, 2, \ldots, K;\ j = 1, 2, \ldots, N_i \right\}$ be a set of training samples for *K* reference actions, each reference action *i* with N_i samples. Training sample $\mathbf{x}_i^j$ indicates the *j*-th template sample of action *i*, and is represented by its EMHI $E_i^j(r, c)$ of size $R \times C$.

The two-dimensional energy image $E_i^j(r, c)$ of training sample $\mathbf{x}_i^j$ is first reshaped as a one-dimensional vector. Denote by $\mathbf{z}_t = [z_t(k)]$ the reshaped one-dimensional vector of the training sample $\mathbf{x}_i^j$ with $t = j + (\sum_{\ell=1}^{i} N_\ell) - N_i$, $i = 1, 2, \ldots, K$, and $j = 1, 2, \ldots, N_i$. The *k*-th element of $\mathbf{z}_t$ is given by

$$z_t(k) = E_i^j(r, c)$$

where $k = c + (r-1) \cdot C$, for $r = 1, 2, \ldots, R$ and $c = 1, 2, \ldots, C$, given that the image is of size $R \times C$.

The de-mixing matrix obtained from the FastICA model is given by $\mathbf{W}$. Therefore, the source action templates can be estimated by

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_T \end{bmatrix} = \mathbf{W} \cdot \mathbf{Z}^{\mathrm{T}}$$

where

$$\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T]\text{, and}$$

$$T = \sum_{i=1}^{K} N_i$$

The data matrices $\mathbf{Z}$ and $\mathbf{U}$ are of size $T \times (R \cdot C)$, and the de-matrix $\mathbf{W}$ is of size $T \times T$. For a total of *T* training samples of the reference actions, we can obtain up to *T* basis images $[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_T]$ from the ICA. Any action $\mathbf{x}$ represented by the EMHI $E(r, c)$ can be constructed by a linear combination of the *T* basis images $\mathbf{u}_i$'s, i.e.,

$$\mathbf{x} = \mathbf{b} \cdot \mathbf{U} = \sum_{i=1}^{T} b_i \cdot \mathbf{u}_i$$

and

$$\mathbf{b} = \mathbf{x} \cdot \mathbf{U}^{+}$$

where $\mathbf{U}^{+}$ is the pseudo inverse of the basis image matrix $\mathbf{U}$, and $\mathbf{U}^{+} = \mathbf{U}^{\mathrm{T}} \left[\mathbf{U} \cdot \mathbf{U}^{\mathrm{T}}\right]^{-1}$. The coefficient vector $\mathbf{b} = (b_1, b_2, \ldots, b_T)$ gives the discriminant features for describing the contents of the action $\mathbf{x}$.

For an unknown testing action $\mathbf{x}_{\text{test}}$ represented by its EMHI, the corresponding discriminant feature $\mathbf{b}_{\mathbf{x}_{\text{test}}}$ is calculated by the trained ICA basis images, i.e.,

$$\mathbf{b}_{\mathbf{x}_{\text{test}}} = \mathbf{x}_{\text{test}} \cdot \mathbf{U}^{+}\ .$$

The distance measure between template sample $\mathbf{x}_i^j$ and the unknown $\mathbf{x}_{\text{test}}$ is defined as

$$\Delta \mathbf{b} = \min \left\| \mathbf{b}_i^j - \mathbf{b}_{\mathbf{x}_{\text{test}}} \right\|, \forall\ \mathbf{b}_i^j\ .$$

where $\mathbf{b}_i^j$ is the coeeficient vector of training sample $\mathbf{x}_i^j$.

Given that the best coefficient vector $\mathbf{b}_{i^*}^{j^*}$

achieves the minimum distance, i.e.,

$$\mathbf{b}_{i^*}^{j^*} = \arg\min_{\mathbf{b}_i^j} \left\| \mathbf{b}_i^j - \mathbf{b}_{\mathbf{x}_{\text{test}}} \right\|$$

The proposed classification then assigns the unknown $\mathbf{x}_{\text{test}}$ to the reference action i^*. In order to rule out a novel action (i.e., an action irrelevant to any of the reference actions of interest), a simple distance threshold can be applied to exclude the unrecognizable action.

3 EXPERIMENTAL RESULTS

This section evaluates the effectiveness of the proposed action recognition scheme with the Public Weizmann action dataset (Gorelick et al., 2007). In the Weizmann dataset, we compare the experimental results of the proposed method and several existing methods with the Weizmann dataset. The Weizmann dataset contains 90 low-resolution (180 × 144) video sequences that show 9 different people, each performing 10 natural actions of Run, Walk, Skip, Jumping-jack, Jump forward on two legs, Jump in place on two legs, Gallop sideways, Wave two hands, wave one hand, and Bend. All the backgrounds are stationary. We use 9 training samples for each individual reference action, wherein each EMHI training sample is obtained from the last frame of each sequence. In the training stage, a total of 90 (10 actions × 9 people) video representation sequences are collected as the training data matrix.

The proposed method gave the same detection results as those of the state-of-the-art methods (e.g., volume-based method [27], spatiotemporal methods (Gorelick et al., 2007); (Hsiao et al., 2008), model-based method [29] and optical flow-based method [30]), as seen in Table 1. The existing methods generally use shapes or silhouettes for the representation, highly rely on accurate segmentation of foreground shapes, and require high computation complexity. They may fail to recognize the actions of interest on a disturbed background.

The proposed algorithms were implemented using the C++ language on a Core2 Duo, 2.53GHz personal computer. The test images in the experiments were 200×150 pixels wide with 8-bit gray levels. The total computation time from foreground segmentation and spatiotemporal representation to ICA feature extraction and distance measure for an input image was only 0.015 seconds. It achieved a mean of 67 fps (frames per second) for real-time action recognition.

Table 1: Performance comparison of different methods on the Weizmann dataset.

Methods	Accuracy
Grundmann *et al.* 2008	94.6 %
Hsiao *et al.* 2008	96.7 %
Wang and Suter 2007	97.8 %
Fathi and Mori 2008	100 %
Gorelick *et al.* 2007	100 %
Our proposed method	100 %

4 CONCLUSIONS

The proposed ICA-based feature extraction and classification method has been well applied to the global spatiotemporal representation of EMHI for recognizing activities that can be observed from a global, macro viewpoint. It is worth further investigation to extend the ICA-based scheme for recognizing subtle activities with micro-observation representation that can be observed only from a local viewpoint of detailed body movements of individual foreground objects.

REFERENCES

Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R., 2007. Actions as Space-Time Shapes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247-2253.

Hyvarinen, A., and Oja, E., 1997. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, vol. 9, pp. 1483-1492.

Hurri, J., Gavert, H., Sarela, J., and Hyvarinen, A., 2004. FastICA Package. Online Available: http://www.cis.hut.fi/projects/ica/fastica/.

Grundmann, M., Meier, F., and Essa, I., 2008. 3D shape context and distance transform for action recognition. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Hsiao, P. C., Chen, C. S., and Chang, L. W., 2008. Human action recognition using temporal-state shape contexts. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, L. and Suter, D., 2007. Recognition human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Fathi, A. and Mori, G., 2008. Action Recognition by Learning Mid-level Motion Features. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.