A real-time ICA-based activity recognition in video sequences

Du-Ming Tsai
Yuan-Ze University
135 Yuan-Tung Road, Chung-Li, Taiwan
iedmtsai@saturn.yzu.edu.tw

Wei-Yao Chiu
Yuan-Ze University
135 Yuan-Tung Road, Chung-Li, Taiwan
S968902@mail.yzu.edu.tw

Abstract

Action recognition has become more and more important over the past years. The currently existing action recognition algorithms have proven successful for the recognition of sign language, aerobics, tennis sports, etc. However, most of the algorithms require a simple and steady background. They may fail when both foreground and background objects present in the video sequence are moving. In this paper, we propose an independent component analysis (ICA) based scheme 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. If the input video contains an action similar to the training sample, then the corresponding EMHI can be reconstructed from the linear combination of the ICA basis images. The coefficients of the linear combination are thus used as the discriminant feature vector for action classification. Compared to the existing methods, the proposed scheme is simple, and yet very effective and computationally very fast for activity recognition. It achieves 67 fps for images of size 200×150. The proposed method does not require the feature design or modeling process. It can be easily implemented for on-line, real-time applications on a low-cost/low-end personal computer or even a portable smart phone. Experimental results reveal that it is robust under disturbed backgrounds where both foreground and background objects are moving simultaneously in the scene. The experiments also demonstrate the potential applications of the proposed method for human-computer interaction and video retrieval.

1. Introduction

Action recognition has become important over the past years. There are many potential applications, including video surveillance, human-computer interaction, gesture recognition, and video retrieval. In conventional action recognition algorithms, the first processing step is foreground segmentation. The foreground person can be segmented using background subtraction or temporal differencing. The second step is spatiotemporal representation from a chunk of video sequences. The third step is the extraction of discriminant features from the motion representation. A good spatiotemporal representation can produce distinct features for robust action classification. Finally, the last step is to classify the actions based on the extracted features.

In this paper, we propose a fast action recognition scheme based on independent component analysis (ICA). 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 bases of action templates from the training EMHI images. Each action to be recognized is then 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.

The Euclidean distance of feature vectors between training actions and testing actions is used as the similarity measure. If the input scene currently under testing resembles a pre-trained reference action, the measured distance should be very small, and all other irrelevant actions should result in large distances. The proposed method does not require the feature designs or modeling processes of individual actions. It needs only a few action samples for the training and is robust under disturbed backgrounds, where 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

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.

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 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 motion history image (EMHI) up to frame t is defined as

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

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

$$M_t(x, y) = \begin{cases} T_{energy}\ , \text{ if } f_t(x, y) \in Foreground \\ 0, \text{ otherwise} \end{cases} \quad (2)$$

and T_{energy} is a energy constant, and is assigned to each foreground pixel. The choice of T_{energy} value is not critical at all as long as it is larger than zero for foreground points and equal to zero for background points. The value of T_{energy} affects only the visual representation of the energy map in the image. It does not change the detection results. In order to avoid digital quantization error and show visually the EMHI in an 8-bit gray-level image, the constant T_{energy} is set to 30 in this paper.

To compare the effectiveness of spatiotemporal representations from the proposed Exponential MHI (EMHI) and the conventional MHI [1], Figure 1 displays the resulting energy images for a set of various actions. Figures 1(a1)~(a3) involve repetitive actions with horizontal movements in the scene. The conventional MHI representations (Figure 1(b1)~(b3)) are too similar to distinguish the different actions of Run, Walk and Side-walk. EMHI (Figure 1(c1)~(c3)), however, can generate distinguishable representations for the different repetitive actions. As seen in Figures 1(a1)~(a3) for the Run, Walk and Side-walk actions, the EMHI representations show the three actions with different leg-motion patterns. Therefore, EMHI gives a better discriminative spatiotemporal representation for describing various action patterns, especially for those that involve continuous, repetitive motions.

Figures 2(a1)~(a3) show the one-hand wave action on a disturbed background. The objects causing disturbance include pedestrians and motorcycles passing through the background. Figures 2(b1)~(b3) present the conventional MHI images with severe interference from the background objects. Figures 2(c1)~(c3) illustrate the proposed EMHI images 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, they reveal that the proposed EMHI can represent different activities that involve simple or complex motions on a disturbed background.

To construct a classification system for action recognition, the classical approach needs 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 a specific selection criterion. It finally applies a classifier to identify individual actions. When a scene involves both foreground and background moving objects, the feature values extracted from the spatiotemporal representation can distinctly deviate from the training samples that contain no moving background objects.

In this study, we develop a robust action recognition that can be tolerant to moving background objects. In order to prevent the tiresome process of feature design and verification of individual actions, we perform an ICA-based method that finds a set of basis images from predetermined multiple reference actions represented by the proposed EMHIs. The basis images resulting from the independent component analysis are statistically independent of each other. Any given input action for recognition is also first represented by the EMHI. The corresponding EMHI is then reconstructed by the linear combination of the ICA basis images. The coefficients of the combination are used as the discriminant feature vector. The Euclidean distance between each training reference action and the input testing action is then calculated. The unknown action is subsequently assigned to the reference action that gives the minimum distance.

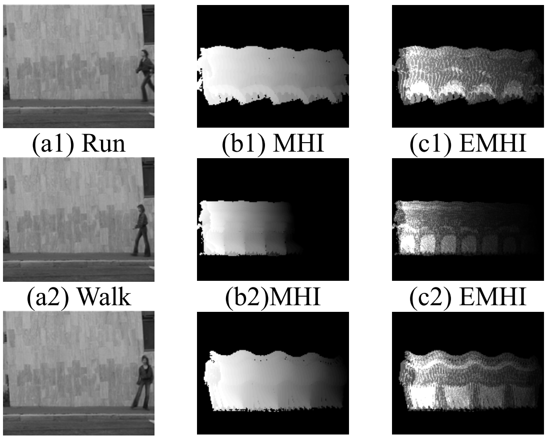

(a1) Run (b1) MHI (c1) EMHI
(a2) Walk (b2)MHI (c2) EMHI
(a3) Side walk (b3) MHI (c3) EMHI

Figure 1. Different activities in the Weizmann dataset and corresponding representations.

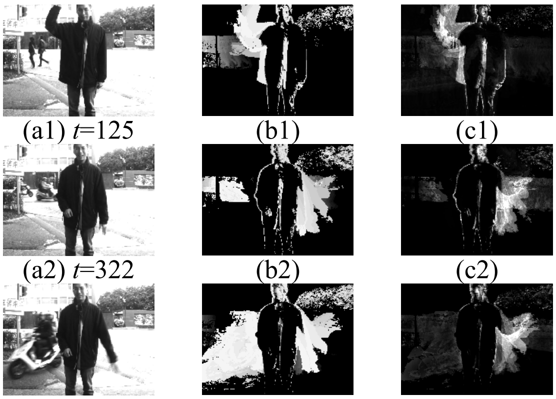

(a1) t=125 (b1) (c1)
(a2) t=322 (b2) (c2)
(a3) t=374 (b3) (c3)
Video sequence MHI EMHI

Figure 2. Repetitive one-hand waving on a disturbed background and corresponding representation. (The symbol t represents the frame number in the video sequence with fps=15.)

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, denoted by $E_i^j(r, c)$, 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

$$z_t(k) = E_i^j(r,\, c) \tag{3}$$

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 **W**. Therefore, the source action templates can be estimated by

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_B \end{bmatrix} = \mathbf{W} \cdot \mathbf{Z}^{\mathrm{T}} \tag{4}$$

where $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_B]$, and $B = \sum_{i=1}^{K} N_i$

The data matrices **Z** and **U** are of size $B \times (R \cdot C)$, and the de-matrix **W** is of size $B \times B$. For a total of *B* training samples of the *K* reference actions, we can obtain up to *B* basis images $[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_B]$ from the ICA. Any action $\mathbf{x}$ represented by the EMHI $E(r,\, c)$ can be constructed by a linear combination of the *B* basis images $\mathbf{u}_i$'s, i.e.,

$$\mathbf{x} = \mathbf{b} \cdot \mathbf{U} = \sum_{i=1}^{B} b_i \cdot \mathbf{u}_i \tag{5}$$

and

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

where $\mathbf{U}^{+}$ is the pseudo-inverse of the basis image matrix **U**, and $\mathbf{U}^{+} = \mathbf{U}^{\mathrm{T}}\left[\mathbf{U} \cdot \mathbf{U}^{\mathrm{T}}\right]^{-1}$. The coefficient vector $\mathbf{b} = (b_1, b_2, \ldots, b_B)$ 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 training sample $\mathbf{x}_i^j$ and the unknown $\mathbf{x}_{\text{test}}$ is defined as

$$\Delta\mathbf{b} = \min\left\|\mathbf{b}_i^j - \mathbf{b}_{\mathbf{x}_{\text{test}}}\right\|,\ \forall\ \mathbf{b}_i^j \tag{6}$$

where $\mathbf{b}_i^j$ is the coefficient vector of training sample $\mathbf{x}_i^j$. Let $\mathbf{b}_{i^*}^{j^*}$ be the best coefficient vector that achieves the minimum distance, i.e.,

$$\mathbf{b}_{i^*}^{j^*} = \arg\min_{\mathbf{b}_i^j}\left\|\mathbf{b}_i^j - \mathbf{b}_{\mathbf{x}_{\text{test}}}\right\| \tag{7}$$

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

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.

The proposed method is used to evaluate the effectiveness on the KTH action dataset [2]. Table 1 compares the accuracy of the proposed method with the state-of-the-art methods [2-5] on the KTH dataset using the same experimental setting. In the comparative methods, Cao et al. [4] provided a good reconstruction rate of 95.02% with same 16 training persons. However, their method with 3D subvolume representation is computationally more intensive and requires more storage space for video sequences. The proposed method collects only 2D EMHI representation for training and yields an accuracy of 98.3% that outperforms other comparative methods on the KTH dataset.

Table 1. Performance comparison of different methods on the KTH dataset.

Methods	Accuracy
Schuldt *et al.* [2]	71.71 %
Yuan et al. [3]	93.30 %
Liu and Shah [5]	94.16 %
Cao *et al.* [4]	95.02 %
Our proposed method	**98.30 %**

We also demonstrate the potential application of the proposed method for video retrieval. The BEHAVE dataset [6], created by the University of Edinburgh, was used for testing. It involves nine different activities, such as Walk Together and Run Together in the image sequences. The BEHAVE dataset has eight video sequences (marked as Sequence 0 to Sequence 7), each containing a different combination of activities. The BEHAVE video images were captured at 25 frames per second. For the first retrieval experiment, three moving-car (Car) events were collected from Sequence 0 for training. The training data contained only three representative EMHIs, as shown in Figure 3. Then all eight sequences were used for testing Car retrieval in the video sequences. The moving cars for testing differed from the training templates in shape and direction of movement. The energy update rate γ was set at 0.99 due to the rapid moving speed of cars. Figure 4 demonstrates the car retrieval results of Sequence 1, where the x-axis is the frame number and the y-axis is the corresponding Euclidean distance $\Delta\mathbf{b}$ of each frame. Besides the moving-car events, Sequence 1 also contained the events of Walk-together, Approach, Ignore and Split. The results show that all moving-car events presented very small distance values in the video sequence. By setting the distance threshold at 20, all target events were detected, and only one false alarm was generated. The false-alarmed event was a group-meeting of 5 people. It continued only for a few frames.

In the second retrieval experiment, three fighting events were collected from Sequence 4, as seen in Figure 5. Figure 6 demonstrates the fighting retrieval results from Sequence 5. In addition to the fighting event, Sequence 5 also contained the events of Group-meeting, Approach, Walk Together, and Split. The fighting event could be detected by a simple threshold. In Sequence 5, the fighting events involved short and long durations and fighting with chasing, which were quite different from the training samples. The fighting events in the test also occurred in different locations in the image scene. They were all quite different from the training samples, yet the proposed method was still able to generate low distance measures for all fighting events with few false alarms.

For video retrieval applications, the detection threshold can be replaced with a desired precision-recall rate. Precision is the percentage of retrieved instances that are relevant, while recall is the percentage of relevant instances that are retrieved. High recall means that an algorithm returns most of the relevant results. High precision means that an algorithm returned more relevant results than irrelevant events.

Figures 7 shows the precision-recall curve of the proposed method for car-event retrieval (Fig. 3) and fighting-event retrieval (Fig. 5) in the BEHAVE dataset, by varying the distance threshold. Given a recall rate of 1, the precision rate is still more than 0.92 for both Car and Fighting scenarios. In Figures 4 and 6, the dashed lines show the thresholds that give recall rates of 1.

4. Conclusions

This paper has presented ICA-based basis images for motion feature extraction from the Exponential Motion History Image (EMHI) for action recognition in a video sequence. It is model-free, and can be easily applied to real-time activity recognition without large training data. The proposed method in its current form cannot recognize two or more different activities occurred in the same scene. When a scene contains two or more distinct activities, the proposed method needs to segment individual activities from the EMHI and then recognize each activity separately. The region-growing technique can be first applied to the EMHI image to segment meaningful regions by starting from the seeds with local maximum energies. It is worth further investigation.

References

[1] J. Davis, A. Bobick, The representation and recognition of human movement using temporal templates, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1997, pp. 928-934.

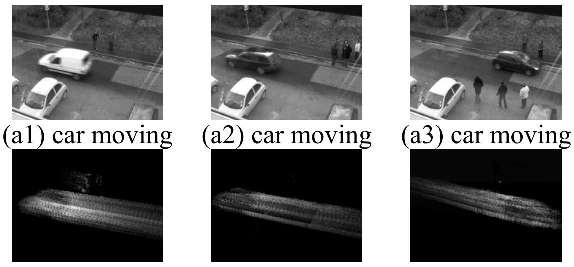
[2] C. Schuldt, I. Laptev, and B. Caputo, Recognizing human actions: A local SVM approach, in: Proceedings of the International Conference on Pattern Recognition, vol. 3, 2004, pp. 32-36.

[3] J. Yuan, Z. Liu, and Y. Wu, Discriminative subvolume search for efficient action detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2442-2449.

[4] L. Cao, Z. Liu, and T. S. Huang, Cross-Dataset Action Detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1998-2005.

[5] J. Liu and M. Shah, Learning human actions via information maximization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1-8.

[6] BEHAVE Dataset, Univ. of Edinburgh, available online: http://groups.inf.ed.ac.uk/vision/behavedata/interactions


(a1) car moving (a2) car moving (a3) car moving

(b1) EMHI (b2) EMHI (b3) EMHI

Figure 3. Training moving-car events in Sequence 0 for video retrieval: (a1)-(a3) a car moving on a street; (b1)-(b3) corresponding EMHIs.

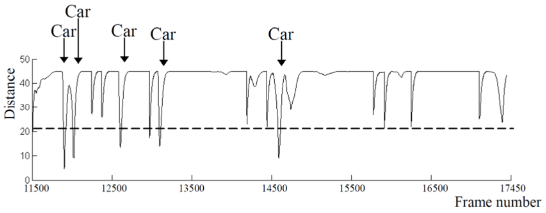
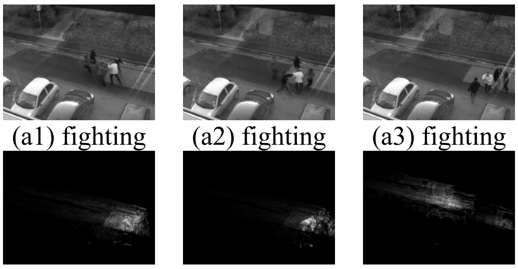


Figure 4. Results of moving-car detection in Sequence 1 of the BEHAVE dataset. (The dashed line is the threshold that gives a recall rate of 1)


(a1) fighting (a2) fighting (a3) fighting

(b1) EMHI (b2) EMHI (b3) EMHI

Figure 5. Training fighting events in Sequence 4 for video retrieval: (a1)-(a3) group fighting and chasing on the street; (b1)-(b3) corresponding EMHIs.

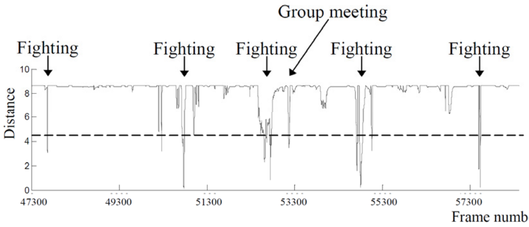
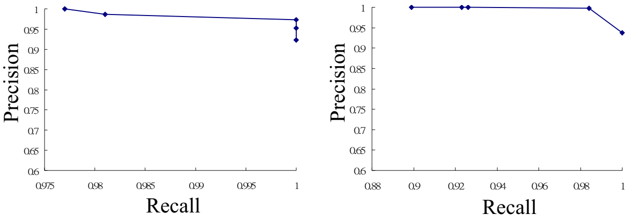


Figure 6. Results of fighting event detection in Sequence 5 of the BEHAVE dataset. (The dashed line is the threshold that gives a recall rate of 1)



(a) Car event (b) Fighting event

Figure 7. Precision-recall curves for event retrieval of the BEHAVE dataset.