MVA '96 IAPR Workshop on Machine Vision Applications, November. 12-14, 1996, Tokyo, Japan

Event Driven Motion-Image Classification by Selective Attention Model

Toshikazu Wada * Takekazu Kato *
Department of Information Technology
OKAYAMA University

Abstract

A motion-image classification method is presented. Our method is designed based on selective attention model which dynamically changes its focusing regions (domains of feature extraction) according to its state so that the essential features can be extracted from input images. The advantages of our method are 1) the feature extraction is not affected by the image variations outside of the focusing regions, 2) feature extraction and the state transition can be computed much faster than other methods, 3) focusing-region sequence can be learned incrementally from training samples , 4) the classifier can be derived from motion-image identifiers trained independently, and 5) the classifier does not output unique result but possible candidates when the input is ambiguous.

1 Introduction

Motion-image classification technology is essential for recognizing the dynamic scene and plays an important role in visual surveillance systems, human interface systems, etc. .

In general, the technology consists of 1) image feature extraction for each image frame and 2) feature sequence analysis. Since the sequence analysis depends on feature extraction which can be affected by the noise and unexpected inputs, bottom-up motion-image classifier cannot be robust. If the image feature extraction can exploit the feature sequence information, a classifier having top-down as well as bottom-up processes can be realized. This will improve the stability of the feature extraction.

Most of the view based motion-image classification methods [1],[2],[3],[4] are designed based on Hidden Markov Model (HMM): a kind of probabilistic automaton. But, the HMM's states are *hidden*, and hence, the feature extraction mechanism referring the current state cannot be embedded. This means that the top-down feature extraction cannot be realized with HMM.

In this paper, we propose a motion-image classifier assuming that each motion-image class is characterized by an event sequence, where the *event* is an image feature at certain time and location. The domain of an event is called *focusing region.*

*Address: 3-1-1, Tsushima-NAKA, Okayama, 700 JAPAN. E-mail: {twada,kato}@chino.it.okayama-u.ac.jp

An example of the event information can be obtained by the background subtraction. In this case, if anomalous pixels fill the focusing regions, the event is detected (=1), otherwise, not(=0). Note that the event descriptions are not affected by the image variations outside of the focusing regions.

If a focusing-region sequence specifying a motion image is known a priori and an event is detected at a certain time, successive events can be obtained by changing the focusing regions.

The advantages of this idea are 1) *the event description is not affected by the outliers* and 2) *feature extraction in the focusing region is much faster than usual one.* However, the feature extractor may lose track of the event sequence only by this idea. To solve this problem, we will introduce multiple interpretations of an event description.

The motion-image classifier is composed of independent motion-image identifiers. In section 2, the identifier based on selective attention model is described, the classifier is described in section 3, and some experimental results are shown in section 4.

2 Motion-image identifier

A motion-image identifier is a simple classifier which accepts motion images in a specified class and rejects those in the other classes. Here we introduce an identifier based on selective attention model.

2.1 Selective attention model

Selective attention model is a successive event detection scheme. The identifier based on this model is driven by events in input images. Since an event is detected at a focusing region, the region must be dynamically changed so as not to lose track of the event sequence.

For the discussion below, the following definitions are given:

Definition 1 (Motion image) *The domain of a motion image is the* spatio-temporal space*: $T \times X \times Y$, where T represents the time axis and $X \times Y$ the image space. In this space, a motion image can be represented as a mapping $I(t,x,y)$: $T \times X \times Y \mapsto P$, where P represents the set of pixel values.*

Definition 2 (Focusing-region sequence) *A focusing-region sequence is the subset of spatio-temporal space, which can be represented as $f(t)$: $T \mapsto \mathcal{B}(X \times Y)$, where $\mathcal{B}(A)$ represents the power set of A. A focusing region is a snapshot of $f(t)$ at a certain time.*

Definition 3 (Event) *An event is a predicate representing the occurrence of an image feature in a focusing region. The event of an image I with a focusing region f can be represented as $e(f, I)$: $\mathcal{B}(X \times Y) \times \mathcal{I} \mapsto \{0, 1\}$, where $\mathcal{I}$ represents a set of images.*

To characterize a motion-image class by an event sequence, event descriptions for motion images in a class are desired to coincide.

Some of the 3-D object motions are restricted by fixed objects or articulated objects having fixed joint, such as wall, door, desk, etc. In this case, common events can be detected by using a focusing region from motion images in a class taken by a fixed camera. But, since apparent motion speeds are different in different motion images, events are detected at different time. That is, time axes are different in different motion images, and hence, the spatio-temporal space is not suitable for the event description.

As for the motion images mentioned above, common event sequences can be detected with a focusing-region sequence by non-linear transformations of the time axes, and the following assumption can be introduced:

Assumption 1 *We assume:*

$$I_i, I_j \in \Omega \Rightarrow \exists \tau_j \exists f \forall t \; e(f(t), I_i(t)) = e(f(\tau_j(t)), I_j(\tau_j(t))), \quad (1)$$

and for such f satisfying the above proposition,

$$I_k \notin \Omega \Rightarrow \forall \tau_k \exists t \; e(f(t), I_i(t)) \neq e(f(\tau_k(t)), I_k(\tau_k(t))) \quad (2)$$

, where Ω represents a motion-image class and τ_j, τ_k are the monotonic functions of t.

In this assumption, the time axis of I_i is regarded as *regularized time* and the function τ_j gives a mapping from the regularized time to the time axis of j-th motion image. Hereafter, we will denote the regularized time q and the inverse transformation of $\tau(q)$ $\rho(t)$. Since both τ and ρ are the monotonic functions, $\rho_i(\tau_i(q)) = q$ and $\tau_i(\rho_i(t)) = t$.

By finding τ or ρ for each motion image, the time variance can be regularized. If the regularized time q and the focusing-region sequence $f(q)$ are given, ρ can be computed as described below:

Since ρ is a monotonic function of t, in a discrete form, the following assumption can be valid under a dense sampling of t and q.

Assumption 2 *At $\rho(t^i) = q^k$, $\rho(t^{i+1})$ must be q^k or q^{k+1} or q^{rej}, where t^i is the i-th sample of t, q^k the k-th sample of the* regularized time, *and q^{rej} the time domain of other motion-image classes.*

Based on this assumption, at $\rho^i = q^k$, ρ^{i+1} can be determined by $e(f(q^k), I^i)$ and $e(f(q^{k+1}), I^i)$, because an event $e(f(q^k), I^i) = 1$ represents the evidence of $\rho^{i+1} = q^k$, and $e(f(q^{k+1}), I_i) = 1$ represents the $\rho^{i+1} = q^{k+1}$, where $I^i = I(t^i)$ and $\rho^i = \rho(t^i)$. The combination of $e(f(q^k), I^i)$ and $e(f(q^{k+1}), I^i)$ is called *event code.*

In this case, the domain of the focusing-region sequence is the space of $Q \times X \times Y$, where Q represents the regularized time, i.e., finite set of *states* defined below. We call this space *event space.*

Table 1: State transition at $\rho^i = q^k$, where q^{rej} represents the image sequence is rejected.

$e(f(q^k), I^i) \cdot e(f(q^{k+1}), I^i)$	ρ^{i+1}
$0 \cdot 0$	q^{rej}
$0 \cdot 1$	q^{k+1}
$1 \cdot 0$	q^k
$1 \cdot 1$	q^k or q^{k+1}

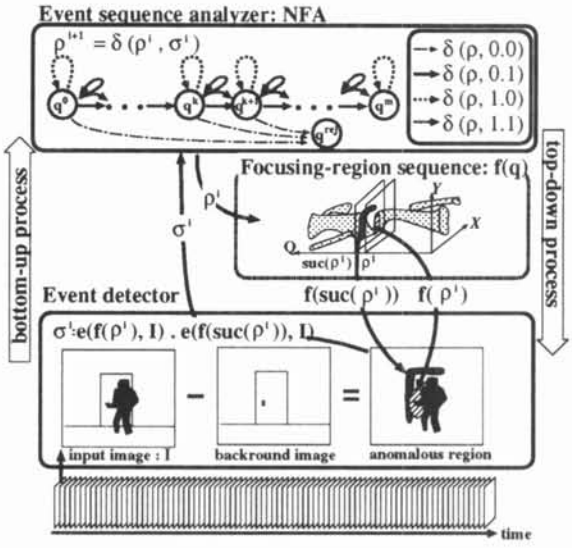


Figure 1: Motion-image identifier

Definition 4 (State) *Q is the ordered set of finite states defined as:*

$$Q = \{q^0, q^1, \cdots, q^m, q^{rej}\}$$

, where $q^i < q^j$ if $i < j$ $(\leq m)$, and $\forall i \; q^i < q^{rej}$. The successor of q^i $(1 \leq i < m)$ is denoted by $suc(q^i)$.

Based on the discussions above, the identifier can be defined as below:

Definition 5 (Motion-image identifier) *Motion-image identifier $M = \{q^0, Q, \delta, e, f\}$ can be defined as*

$$\begin{cases} \rho^0 & = & q^0, \\ \rho^{i+1} & = & \delta(\rho^i, \sigma^i), \\ \sigma^i & = & e(f(\rho^i), I^i) \cdot e(f(suc(\rho^i)), I^i) \end{cases} \quad (3)$$

,where $\rho^i \in Q$ and the state transition δ at $\rho^i = q^k$ is given in Table1.

This model consists of event sequence analyzer and an event detector as shown in Figure1. Since the event detector changes focusing regions according to the current state, we call this model *selective attention model.* Note that this model is not deterministic, because an event code $\sigma^i = 1 \cdot 1$ causes a non-deterministic state transition to q^k and q^{k+1}. This feature makes the model robust.

Formally, this model consists of two major components: a non-deterministic finite automaton (NFA) having the set of event codes [1] $\Sigma = \{0 \cdot$

[1]The event code can be extended as $\sigma^i = e(f(\rho^i), I^i) \cdot e(f(suc(\rho^i)), I^i) \cdot e(f(suc(suc(\rho^i))), I^i) \cdots$, which means the extension of Assumption 2. This extension is effective for recognizing the fast motion images.

$0, 0 \cdot 1, 1 \cdot 0, 1 \cdot 1\}$ as its alphabet which can be represented as: $Q \times \Sigma \mapsto Q$, and the event detector: $\mathcal{B}^2(X \times Y) \times \mathcal{I} \mapsto \Sigma$. The focusing-region sequence $f(q)$ gives a mapping:$Q \mapsto \mathcal{B}(X \times Y)$ which combines these components.

In most of the state transitive motion-image classifiers, the relationship between state and image space is not explicitly hold, and the top-down feature extraction cannot be realized. But, our method enables both bottom-up state transition and top-down feature extraction exploiting this relationship, i.e., the focusing-region sequence $f(q)$ in event space.

2.2 Learning for anomalous-region features

As for the anomalous-region features detected by the background subtraction, the focusing-region sequence can easily be acquired from the anomalous regions of motion-images in a class.

The time axis of anomalous region $a_i(t)$ $(i = 1, 2, \cdots, n)$ can be regularized so as to maximize the following value:

$$\int \frac{|a(q) \cap a_i(\tau_i(q))|}{|a(q) \cup a_i(\tau_i(q))|} dq \tag{4}$$

,where $a(q)$ is the standard sample of the class and $|\cdot|$ represents the number of pixels. This normalization can be done by the dynamic programming (DP).

From the regularized anomalous regions, the focusing-region sequence $f(q)$ in the regularized time q can be computed as:

$$f(q) = \bigcap_{i=1}^{n} a_i(\tau_i(q)). \tag{5}$$

This method enables incremental learning which can be represented as:

$$\begin{aligned} f_1(q) &= a_1(q), \\ f_i(q) &= f_{i-1}(q) \cap a_i(\tau_i(q)) \end{aligned}$$

, where $a_1(q)$ is taken as the standard sample.

This means that the training samples a_i $(i > 1)$ can be abandoned after f_i is computed.

2.3 Event for anomalous-region features

Event can be defined for the anomalous-region features as

$$e(f(q), I_i(t)) = \begin{cases} 1, & f(q) \cap a_i(t) = f(q) \\ 0, & otherwise \end{cases}. \tag{6}$$

This definition satisfies $e(f(q), I_i(\tau_i(q))) = 1$ for training samples. But, in practice, since the training samples do not coincide with the inputs, the event should be defined as

$$e(f(q), I_i(t)) = \begin{cases} 1, & f(q) = \phi \ or \ \frac{|f(q) \cap a_i(t)|}{|f(q)|} > \theta \\ 0, & otherwise \end{cases} \tag{7}$$

, where θ $(0 < \theta \leq 1)$ represents a threshold.

3 Motion-Image Classifier

The classifier consists of independent identifiers $M_{\omega i} = \{q^0_{\omega i}, Q_{\omega i}, \delta_{\omega i}, e, f_{\omega i}\}$ $(i = 1, \cdots, N)$. By introducing an initial state q^0 and ϵ-state transitions[2] from q^0 to $q^0_{\omega i}$, the classifier can easily be realized as shown in Figure2.

[2]The ϵ-state transition means a state transition caused by null input.

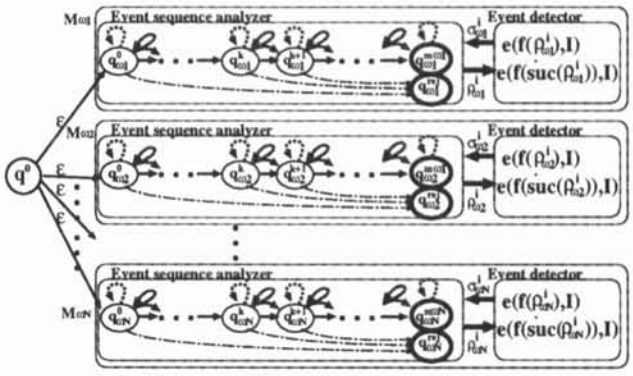
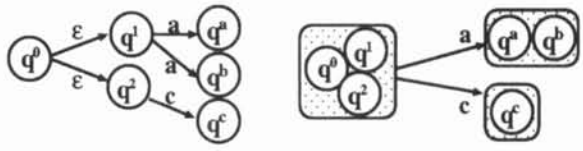


Figure 2: Motion-image classifier



(a) NFA (b) Equivalent DFA

Figure 3: Transformation from NFA to DFA

Since the classifier also has non-deterministic state transitions, it must has multiple current states $\{\rho^k\}$. Inputs are classified according to the contents of $\{\rho^k\}$ as described below:

OnTheWay: Not enough images have been input.
$\exists i \ \ ((q^{m\omega i}_{\omega i} \notin \{\rho^k\}) \wedge (q^{rej}_{\omega i} \notin \{\rho^k\}))$

Rejected: The input doesn't belong to the known classes.
$\forall i \ \ ((q^{m\omega i}_{\omega i} \notin \{\rho^k\}) \wedge (q^{rej}_{\omega i} \in \{\rho^k\}))$

Ambiguous: The input can be classified to multiple classes.
$\exists i \ \ ((q^{m\omega i}_{\omega i} \in \{\rho^k\}) \wedge (\exists j \neq i \ \ (q^{m\omega j}_{\omega j} \in \{\rho^k\})))$

Classified: The input can be classified to ω_i
$\exists i \ \ ((q^{m\omega i}_{\omega i} \in \{\rho^k\}) \wedge (\forall j \neq i \ \ (q^{m\omega j}_{\omega j} \notin \{\rho^k\})))$

3.1 Single process implementation

Since the classifier consists of an NFA and event detectors, it can be implemented by the parallel processing, which can consume much of the computer resources if a lot of non-deterministic state transitions are occurred.

Fortunately, the identifier and the classifier can be transformed into an equivalent deterministic model and can be implemented by a single process as described below:

An NFA can be transformed into an equivalent deterministic finite automaton (DFA) as shown in Figure3. The transformed DFA has subsets of Q as its states. The focusing regions of the event detector can be reorganized by the set operation according to this transformation.

4 Experiments

We implemented the proposed method and applied to the classification of two types of human-motions at a door: "entering" and "exiting" by using anomalous-region features, where the threshold value to detect the anomalous regions is 20.

Figure 4: Motion image example and its anomalous regions of "exiting".

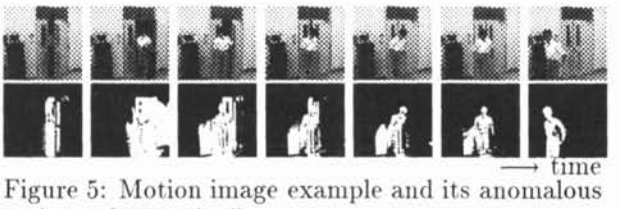


Figure 5: Motion image example and its anomalous regions of "entering".

Figure 6: Focusing regions for "exiting" → state

Figure 7: Focusing regions for "entering" → state

The number of motion-image samples is 16 ("entering":8, "exiting":8) and the samples consist of $167 \sim 319$ frames. The class "exiting" includes motion images where objects appear from both left and right sides of the image frame. Also, the class "entering" includes human motions to left and right sides. Motion-image examples of these classes are shown in Figure 4 and 5.

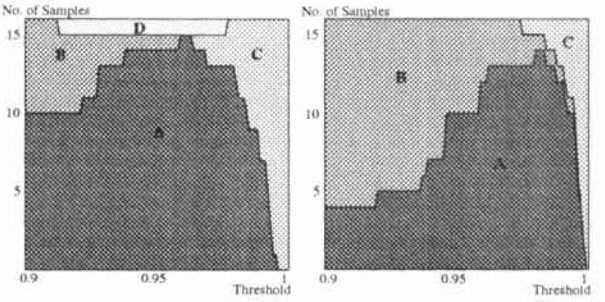
In this experiment, 15 samples are used for learning, and the residual is used as input. In the learning stage, to suppress the meaningless non-deterministic state transition, successive null focusing regions ($= \phi$) at head and tail are cut off. Also, to synchronize the focusing-region sequence with the input, a null focusing region is added to the head of the focusing-region sequence, which causes $e(q^0_{\omega i}, I) = 1$ for any images. This means that current states always include initial states $q^0_{\omega i}$, and hence, initial states are neglected in the classification stage. The focusing-region sequences are shown in Figure 6 and 7.

By shifting the input and training samples, all samples are classified. The classification results for event-code lengths 2 and 3 are shown in Figure8, where the horizontal axes represent the threshold θ, and the vertical axes number of samples.

From this figure, we can notice that the longer event code and lower threshold make the classifier robust but indecisive. But, note that the ambiguous result can be reclassified into "exiting" or "enter" by the other criteria, such as the first come, first served basis. Based on this property of our method, one can decide the suitable threshold value and code length for one's task.

5 Conclusions

We proposed an event driven motion-image classification method integrating bottom-up and top-down processes, which has the following advantages:



Event-code length: 2 Event-code length: 3

Figure 8: Classification Results: A: Correct, B: Ambiguous, C: Rejected, D: Misclassified

- Since focusing regions reduce the domain of feature extraction, event detector is not affected by the outliers.
- For the same reason, real-time motion-image classification can be realized.
- Incremental learning of the focusing-region sequence can be performed for training samples.
- The classifier consists of identifiers which can be trained independently. This enables the classification with a large number of classes.
- The classifier does not produce unique result but possible candidates when the input is ambiguous.

Our method assumes that the moving objects in a class have the similar trajectory in the image space. In spite of this limitation, our method can be used in many surveillance tasks, because in most of the surveillance tasks, object motion is restricted by fixed objects or articulated objects having fixed joint, such as wall, door, desk, ... etc. . However, our method cannot be used for position-free motion classification. This limitation can be removed by using the background subtraction method with pan-tilt-zoom control proposed in [5]. That is, by controling the pan-tilt-zoom parameters so as to normalize the size and location of the anomalous regions in the image space, we can realize the position free motion recognition. This will be done in the future works.

References

[1] Yamamoto J., Ohya J., and Ishii K., "Recognizing human action in time-sequential images using hidden markov model", Proc. of CVPR, pp. 664-665, (1992)

[2] Starner T. and Pentland A., "Real-time american sign language recognition from video using hidden markov models", Proc. of ISCV, pp. 265-270, (1995)

[3] Bregler C. and Omohundro S.M., "Nonlinear manifold learning for visual speech recognition ", Proc. of ICCV, pp.494-499, (1995)

[4] Wilson A. and Bobick A.,"Learing Visual Behavior for Gesture Analysis", M.I.T. Media Laboratory Perceptual Computing Section Technical Report No.337. (1995)

[5] Wada T., Matsuyama T., "Appearance Sphere: Background Model for Pan-Tilt-Zoom Camera ", Proc. of ICPR, vol. A, pp. 718-722, (1996)