Memory Based Human Region Detection

Ayaka Yamamoto, Yoshio Iwai, Hiroshi Ishiguro
Graduate School of Engineering Science, Osaka University
iwai@sys.es.osaka-u.ac.jp

Abstract

Background subtraction is widely used for detecting moving objects; however, color similarity between a background and a moving object is still an important problem. In this paper, we present an exemplar based approach to efficiently detect walking persons that are similar in color to the background. Exemplar database is constructed from couples of degraded and complete human silhouette patterns; the missing parts in an input image obtained by background subtraction are detected by using degraded patterns in the database, and then are compensated by using complete patterns in the database. In our approach, the relationship between the degraded pattern and the complete human silhouette patterns is formulated and compensation of degraded area is performed by optimization in a block-based manner. The experimental results and evaluations of our system are demonstrated.

1 Introduction

Object detection from image sequences has an important role on computer vision because it can be applied to various applications such as person recognition, traffic monitoring and security systems. To realize these applications, many approaches for object detection have been proposed, and brought significant improvements of the detection accuracy.

There are many object detection methods, which are based on various types of features[1], inter-frame difference[2] and background subtraction[3]. Especially, a background subtraction is commonly used for its simpleness. Background models, however, should be sophisticated at pixel region or frame level to deal with dynamic and local illumination changes.

Many statistical approaches have also been proposed[4]. Stauffer et al. use a mixture Gaussian model for changes in background pixel values[4]. To simultaneously address the problems of illumination changes and casting shadows, a linear model is used for expressing a pixel value, which decomposes into two components: ambient and direct light[5]. This method can detect moving objects in real time, but the method assumes that scene is static, so waving leaves are detected in the background as moving objects (see Fig. 1). The intrinsic problem of background subtraction is that objects similar in color to the background cannot be detected. These methods also have the same problem. These two problems are trade-off because if the range of background color becomes larger in order to remove waving leaves, the area of undetected object similar in color to the background becomes larger. Therefore, the prior knowledge should be required to avoid the problems.

As the prior knowledge of object region, geometric constraints of object's features are memorized, and then the whole body of an object is estimated from detected parts of an object[6]. The estimated objects are restricted that can be represented by geometric constraints. In this paper, we use couples of degraded and complete human silhouette patterns as the prior knowledge. By using the pair of degraded and complete patterns, degraded area are compensated by complete patterns. Collins et al. have also proposed a detection method by using complete human silhouette, but they used the whole body silhouette as one exemplar, so the method requires enormous number of exemplars[7]. Our approach, however, is to use degraded and complete patterns in a block-based manner, so less number of exemplars is required to express various human silhouettes than Collins' method. In this paper, the problem of object detection is modeled as an optimization problem, i.e. combination problem, and object regions are detected by minimizing the objective function.

2 Human Region Detection As Optimization Problem

2.1 Exemplar Database

Exemplar database used in this paper consists of pairs of degraded human region pattern (hereafter, we denote this pattern as degraded pattern) and complete human region pattern (complete pattern). A retrieval key is a degraded pattern acquired by background subtraction. The compensation process is performed by replacing degraded patterns with complete patterns retrieved from the exemplar database.

Let $\boldsymbol{M}$ and $\boldsymbol{C}$ be a binarized human region (image) obtained by background subtraction and a binarized human region (image) obtained by hand, respectively. Both images are divided into overlapped blocks $w \in \mathcal{W}$ of B×B pixels. A site $w^* \in \mathcal{W}^* = \mathcal{I} \times \mathcal{W}$ is defined by a serial number of an input image, $i \in \mathcal{I}$, and a location of a block, w. A pair of an image patch $\boldsymbol{M}_{w^*}$ at site w^* in a degraded image and an image patch $\boldsymbol{C}_{w^*}$ at site w^* in a complete image is registered to a database as an exemplar. The process flow of constructing an exemplar database is shown in Fig. 2.

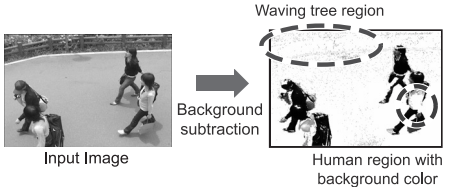

Figure 1. Problems of the previous method[5]

2.2 Formulation

In order to estimate complete human regions by using the exemplar database, some criteria are required to select the best complete pattern from a degraded pattern. For this purpose, we define an objective function to select a complete pattern from a degraded pattern.

Let $\boldsymbol{I}$ be a binarized image obtained by subtracting a background image from an input image. The binarized image $\boldsymbol{I}$ is divided into blocks $v \in \mathcal{V}$ of B × B pixels. The problem is to find a complete pattern $\boldsymbol{C}_{w^*}$ at each block v from an input pattern $\boldsymbol{I}_v$ in order to generate an estimated image $\hat{\boldsymbol{I}}$. The objective function that evaluates the completeness of an estimated image $\hat{\boldsymbol{I}}$ is defined as follows:

$$E(\boldsymbol{I}) = \sum_{v \in \mathcal{V}} \left[D(\boldsymbol{I}_v, \boldsymbol{M}_{w^*}) + \kappa \sum_{u \in \mathcal{N}(v)} D(\hat{\boldsymbol{I}}_u, \boldsymbol{C}_{u^*}) \right], \quad (1)$$

where κ is a constant, $u \in \mathcal{N}(v)$ is a neighbor block of v, $u^* \in \mathcal{N}(w^*) = \{(i,u); u \in \mathcal{N}(w)\}$ is a neighbor block of site w^*. Here, neighborhood $\mathcal{N}(v)$ does not contain v. $D(\cdot,\cdot)(0 \le D \le 1)$ expresses a distance between image patterns. In this paper, we use Hamming distance normalized by block size as the distance function $D(\cdot,\cdot)$. The first term in equation (1) is a data term in proportion to a distance between an input pattern $\boldsymbol{I}_v$ and a degraded pattern $\boldsymbol{M}_{w^*}$. Minimizing the data term is equivalent to find a pattern similarly degraded to an input pattern. The second term is a smooth term in proportion to a distance between an estimated pattern $\hat{\boldsymbol{I}}_u$ neighbor to v and a complete pattern $\boldsymbol{C}_{u^*}$. Minimizing the smooth term is equivalent to find a pattern most likely to fit from neighborhood patterns $\boldsymbol{C}_{u^*}$. Figure 3 shows the relation among input, degraded, complete and estimated patterns.

When an effect of smoothness is strong, we can get smooth human regions, but peak areas such as foot, and hand cannot be compensated. Therefore, we introduce a penalty term to preserve foot and hand area from smoothing at boundaries as follows:

$$E(\boldsymbol{I}) = \sum_{v \in \mathcal{V}} \left[D(\boldsymbol{I}_v, \boldsymbol{M}_{w^*}) + \kappa \sum_{u \in \mathcal{N}(v)} D(\hat{\boldsymbol{I}}_u, \boldsymbol{C}_{u^*}) + g(\boldsymbol{I}_v, \boldsymbol{C}_{w^*}) \right],$$

$$g(\boldsymbol{I}_v, \boldsymbol{C}_{w^*}) = \lambda \frac{|\boldsymbol{I}_v|}{B^2} \quad \text{if } \left| {}^{\exists}\hat{\boldsymbol{I}}_u \right| = 0, u \in \mathcal{N}(v) \text{ and } |\boldsymbol{C}_{w^*}| = 0, \quad (2)$$

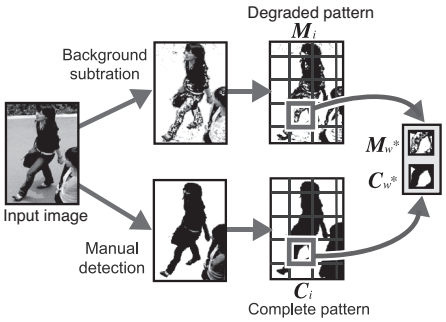

Figure 2. Exemplar creation process

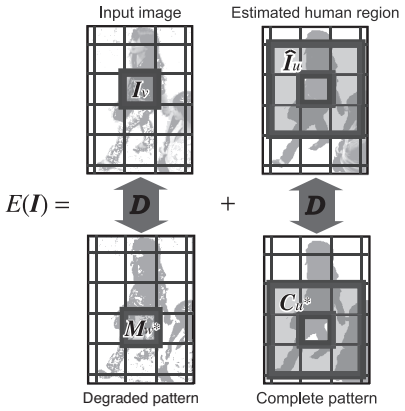

Figure 3. Objective function

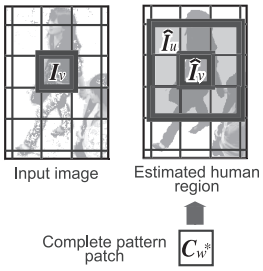

Figure 4. Relationship between patches in the penalty term

where λ is a constant, $u \in \mathcal{N}(v)$ is a neighbor block to v, and $|\cdot|$ expresses the number of pixels of a human region. When at least one neighbor block $\hat{\boldsymbol{I}}_u$ does not contain a human region, v is regarded as a boundary block. When a complete pattern $\boldsymbol{C}_{w^*}$ is assigned to a boundary block, the penalty term is added to the objective function in proportion to the number of pixels of a human region. Therefore, it is avoided that a complete pattern without a human region is assigned to a boundary block. Figure 4 shows the relation between a penalty term and patterns.

3 Exemplar Based Human Region Detection

In this paper, we use binarized images by subtracting input images from background images obtained by the adaptive background model[5]. An estimated complete image $\hat{\boldsymbol{I}}$ is initially generated from an input image $\boldsymbol{I}$, and then minimize the objective function E at each block v by retrieving complete patterns from the exemplar database. When all blocks are replaced by complete patterns $\boldsymbol{C}_{w^*}$, the objective function $E(\hat{\boldsymbol{I}})$ is recalculated for iteration. The minimization loop is finished when $E(\hat{\boldsymbol{I}})$ does not change or the number of iteration is larger than a certain threshold.

3.1 Initialization of estimated image

At each block v, a degraded pattern $\hat{\boldsymbol{M}}_{w^*}$ is selected that is the nearest pattern to an input pattern $\boldsymbol{I}_v$ as

Table 1. Comparison of the number of un-/over-detected pixels and evaluation values with/without penalties

	penalty	human region	undetected region	over-detected region
(1) B = 28	off	21609	1573	1837
	on	21561	1601	2145
(2) B = 32	off	18542	3855	1052
	on	19941	2520	1116
(3) B = 36	off	20932	2436	2023
	on	20932	2436	2023
(4) B = 40	off	20018	3111	1784
	on	20018	3111	1784

follows:

$$\hat{\boldsymbol{M}}_{w^*} = \arg\min_{w^*} D(\boldsymbol{I}_v, \boldsymbol{M}_{w^*}). \quad (3)$$

The initial estimated image is generated from a complete pattern $\boldsymbol{C}_{w^*}$ at each block v corresponding to the degraded pattern $\hat{\boldsymbol{M}}_{w^*}$ in the exemplar database. When there are multiple patterns nearest to an input pattern, the complete pattern that mostly contains human regions is selected in order to preserve a human region.

3.2 Update of estimated image

Once an initial estimated image is obtained, assignment of complete pattern $\boldsymbol{C}_{w^*}$ at each block v is updated in order to minimize the objective function. At each block v an exemplar is retrieved that minimizes the following equation:

$$D(\boldsymbol{I}_v, \boldsymbol{M}_{w^*}) + \kappa \sum_{u \in \mathcal{N}(v)} D(\hat{\boldsymbol{I}}_u, \boldsymbol{C}_{u^*}) + \lambda \frac{|\boldsymbol{I}_v|}{B^2}, \quad (4)$$

and then an estimated pattern at block v is replaced by a complete pattern corresponding the retrieved exemplar. When there are multiple exemplars, an exemplar is chosen that the second term in equation 4 is the smallest.

4 Experimental Results

We conducted evaluation experiments to verify the efficiency of the proposed method in the view points of penalty term and block size. In this experiment two image sequences are used. One sequence (scene 1) is a scene of a sunny and windy day, and leaves are waving by wind (see Fig. 5). Human regions extracted by the Adaptive Background Model[5] are often over-detected in tree regions (see Fig. 6 top). Persons are walking horizontally and human regions are almost same size. The other sequence (scene 2) is also a scene of sunny and windy day (see Fig. 7), and persons are walking vertically and the sizes of human regions are variously changed as shown in Fig. 8. An exemplar database is made from 170 frames in scene 1 and 45 frames in scene 2, and test images are chosen from frames that the exemplar database does not contain. The size of binary image is 720 by 486 pixels.

Figure 9 shows input images, manually detected human images, and estimated images with/without the penalty term by changing block sizes: 28, 32, 36, 40 pixels. The number of pixels of human regions is 21345 pixels. The size of undetected human regions in an input image is 3285 pixels and 3360 pixels in background are misclassified as a human region. Table 1 shows the number of pixels of a human region, over-detected human region, undetected human region.

From Table 1, the number of pixels of undetected human regions are reduced by adding the penalty term, and this reduction can be seen in the bottom left image in Fig. 9. From Fig. 9, a foot pattern is sometimes placed at opposite direction because the proposed method only considers the local blocks of the target block and does not consider the whole body of human. It is a trade-off problem if walking persons are occluded each other.

5 Conclusion

In this paper, we proposed an exemplar based approach to efficiently detect complete human regions from a degraded image acquired by a subtraction method. The proposed method can complement various degrade regions from finite couples of degraded and complete human regions as exemplar data. In order to estimate complete human regions, we introduced an objective function between an input image and exemplar data. We showed through experimental results that complete human regions can be obtained by using the objective function and the best block size can be selected from the proposed criterion. In future work we plan to adopt the proposed method by fast retrieving algorithm from the exemplar database because the optimization takes much time.

References

[1] Zhang, L., Nevatia, R.: Efficient scan-window based object detection using GPGPU. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2008) 1–7

[2] Huang, K., Wang, L., Tan, T., Maybank, S.: A real-time object detecting and tracking system for outdoor night surveillance. PR **4** (2008) 432–444

[3] Fukui, S., Iwahori, Y., Woodham, R.J.: GPU based extraction of moving objects without shadows under intensity changes. In: IEEE Congress on Evolutionary Computation. (2008) 4165–4172

[4] C.Stauffer, Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, USA (1999) 246–252

[5] Yamamoto, A., Iwai, Y.: Real-time object detection with adaptive background model and margined sign correlation. In: the 9th Asian Conference on Computer Vision, Xi'an, China (2009) 65–74

[6] Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: ECCV workshop on statistical learning in computer vision. (2004) 17–32

[7] Collins, R., Gross, R., Shi, J.: Silhouette-based human identification from body shape and gait. In: 5th Intl. Conf. on Automatic Face and Gesture Recognition. (2002)

Figure 5. Scene background

Figure 6. Examples of exemplars; (top) degraded pattern images (bottom) complete pattern images

Figure 7. Scene background

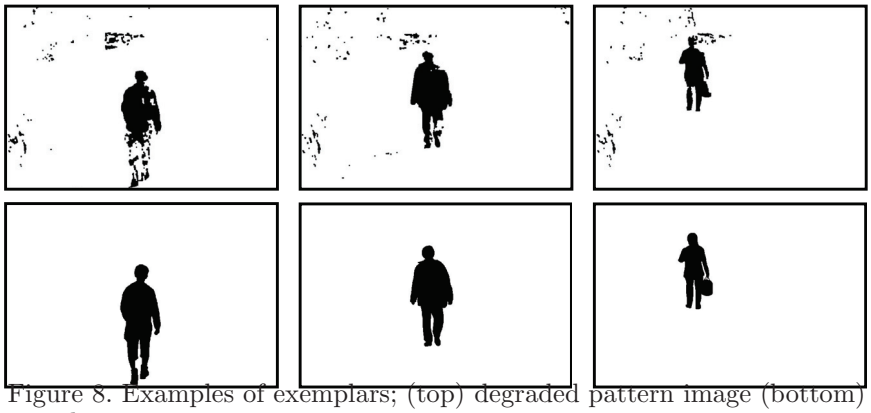

Figure 8. Examples of exemplars; (top) degraded pattern image (bottom) complete pattern image

(top) input image
(btm.) manual detection

(1)B = 28 (2)B = 32 (3)B = 36 (3)B = 40

Figure 9. Comparison of estimation results; (top) without penalties (bottom) with penalties