SUMMARY  We propose a novel motion segmentation method based on a Clausius Normalized Field (CNF), a probabilistic model for treating time-varying imagery, which estimates entropy variations by observing the entropy definitions of Clausius and Boltzmann. As pixels of an image are viewed as a state of lattice-like molecules in a thermodynamic system, estimating entropy variations of pixels is the same as estimating their degrees of disorder. A greater increase in entropy means that a pixel has a higher chance of belonging to moving objects rather than to the background, because of its higher disorder. In addition to these homologous operations, a CNF naturally takes into consideration both spatial and temporal information to avoid local maxima, which substantially improves the accuracy of motion segmentation. Our motion segmentation system using CNF clearly separates moving objects from their backgrounds. It also effectively eliminates noise to a level achieved when refined post-processing steps are applied to the results of general motion segmentations. It requires less computational power than other random fields and generates automatically normalized outputs without additional post-processes.

key words: motion segmentation, pixel classification, time-varying imagery

1. Introduction

The goal of unsupervised motion segmentation is to extract moving objects in the field of view of stationary cameras. It is an essential low-level step used in many visual surveillance, object detection, object recognition, and object tracking applications, as well as in semantic processes and the latest MPEG video standards [1]–[4]. Motion segmentation generates powerful cues for image understanding and contributes toward improving the effectiveness of motion analyses by reducing the amount of data to be processed [5].

Many methods have been devised for accurate motion segmentation utilizing optical flow techniques [6], [7], motion discontinuities [8], [9], global parametric motion models [10], [11], or background removal techniques [1], [12], [13]. Although various approaches to motion segmentation have been reported, there is a wide consensus that an ideal and polished motion segmentation algorithm should do the following:

- segment only moving objects from the background clearly;
- be robust to noise produced by raw input image sequences;
- be adaptive to time or illumination changes in order to

be robust to background changes by sunshine or variation of light sources;
- appropriately, as part of the background, such repeated motions as waving leaves, reflections on water surfaces, or monitor flickers; and
- appropriately treat an object that has moved, but has remained stopped for a long time, as a provisional part of the background.

To meet these requirements, Stauffer [1] adopted a Gaussian mixture model (GMM) and maintained the Gaussian distributions in each pixel. However, because moving objects are represented as gradually moving groups of pixels in consecutive images with an elapsed time, we should utilize both spatial and temporal information by considering neighbor pixels and previous frames together. Migdal [12] adopted Markov random fields (MRFs) [14] in order to consider both spatial and temporal information, and demonstrated comparatively good motion segmentation results. However, for random fields such as MRFs or conditional random fields (CRFs), it is necessary to estimate some values using simulated annealing (SA), improved iterative scaling (IIS) [15], or a Gibbs sampler [14]. These are iterative processes that require a lot of computational power, and thus it is difficult to process high-resolution images in real time for general systems. Moreover, most random fields used in motion segmentation are used as post-processing steps to reduce noises resulting from motion segmentation. Fazli et al. [16] also applied a post-processing step to GMM-based quantized results for the same purpose. However, quantization necessarily accompanies a reduction of information, which may generate incorrect results, as the reduced information cannot be successfully conveyed in the post-processing steps. Thus, we need a strategy that can reduce the number of quantizing steps.

We adopt a thermodynamical approach, a Clausius normalized field (CNF) [19], to build motion segmentation systems that meet the requirements noted above and to solve the problems mentioned in the previous paragraph. As mentioned in earlier work [19], the essential concept of CNF is derived from the Clausius entropy definition [17] and a thermodynamic system that exchanges heat with its surroundings. Information about well-known four laws of thermodynamics [30] can help understanding this paper well.

To design the CNF, we assume an ideal lattice plane \(P_h\) made up of many cliques of molecules. We also assume that \(P_h\) is hotter than the surrounding atmosphere and

\[E_u = k T \sum_i \log P_i\]

Copyright © 2014 The Institute of Electronics, Information and Communication Engineers
is cyclically contacted by a different plane \( P_a \) hotter than \( P_b \). \( P_b \) absorbs heat from \( P_a \) and emits heat to its surroundings due to differences of temperature among the planes \( P_b \) and \( P_a \) and atmospheres. The heat absorbed from \( P_a \) makes the entropy of \( P_b \) high. Consequently, the cliques of \( P_b \) become more unstable due to an increased disorder [18].

2. Motion Segmentation Using a Simple Clausius Normalized Field

To use CNF for motion segmentation, we need to define the following three features as mentioned in the paper [19]:

- An ideal system that exchanges heat.
- A function to estimate a specific measure, \( dQ \).
- An altering rule of temperatures, \( T(Q) \).

In this paper, to prevent any confusion, most of our mathematical notations are identical in terms of meaning to those in the paper [19]. An ideal system generates \( p'(i, j, t) \) with an update of its parameters along with some physical rules. With a higher variation in entropy, \( p'(i, j, t) \) of a pixel, the pixel is more likely to be classified as a part of the foreground because the instantaneous variation of the disorder of the pixel is also high.

2.1 Defining an Ideal System

We assume an ideal lattice plane, \( P_b \), as shown in Fig. 1, which is made up of many cliques, where each clique corresponds to each pixel in an image. The temperatures of the cliques (pixels) can differ from each other, and each clique (pixel) is comprised of many molecules that absorb heat from input images at every frame. \( P_b \) also emits heat to the surrounding atmosphere at every frame, and molecules with higher temperature states emit more heat. If the absorbed heat is greater than the emitted heat, the temperature, \( T \), increases, and vice versa. According to Clausius entropy definition [17], the amount of entropy variation, \( dS \), is affected by both the temperature, \( T \), and the absorbed heat, \( dQ \).

2.2 Defining \( dQ \)

Let the number of channels of a color space be \( C \), and heat absorbed from an input image be defined as follows:

\[
dQ_{i,j,t} = n \sum_{c} w_c \cdot (X_{c,i,j,t} - b_{c,i,j,t})^2,
\]

where \( n \) is the number of molecules of a clique (of a pixel), \( X_{c,i,j,t} \) is a pixel value (intensity of each channel) of the input image, \( b_{c,i,j,t} \) is a pixel value of \( P_b \), and \( w_c \) is the weight of each channel. We use YCrCb rather than RGB, because while the meanings of the levels of each channel of the RGB color space are equivalent, the meanings of the levels of each channel of the YCrCb color space differ from each other. The effect of \( w_c \) is described in the experiment section. In a real physical system, when different objects meet each other, the greater the difference of temperature between the objects, the greater the heat that is exchanged during a unit period of time. We similarly define \( dQ_{i,j,t} \) as large differences between \( b_{c,i,j,t} \) and \( X_{c,i,j,t} \) that make a large \( dQ_{i,j,t} \).

In order to allow the motion segmentation system to be adapted to an elapse of time, \( b_{c,i,j,t} \) is updated every frame as follows:

\[
b_{c,i,j,t} = (1 - \lambda) \cdot b_{c,i,j,t-1} + \lambda \cdot X_{c,i,j,t},
\]

where \( \lambda \) is the updating rate and \( t \) is the time or frame index. A large value of \( \lambda \) allows the system to adapt to new pixel values in a short period of time.

2.3 Defining \( T(Q) \)

The absolute temperature of a clique (pixel) is proportional to the average energy of the molecules that comprise the clique (pixel) [20]. Thus, if we let the rate constant between absorbed heat per molecule and temperature be \( \kappa \), the amount of variation of temperature by the absorbed heat is

\[
\frac{dT_{i,j,t}}{n} = \kappa \cdot \frac{dQ_{i,j,t}}{n}.
\]

The altering rule of temperatures, \( T(Q) \), is also defined as follows:

\[
T_{i,j,t} = (1 - \rho) \cdot T_{i,j,t-1} + \rho \cdot dT_{i,j,t},
\]

where \( \rho \) is a constant value related to naturally emitted heat. According to (4), the amount of emitted heat per frame is

\[
dQ_{i,j,t}^E = \frac{\sigma \cdot n \cdot \rho T_{i,j,t}}{\kappa},
\]

where \( \sigma \) is a rate constant. This means that \( T(Q) \) obeys the laws of nature in that molecules under higher temperature emit more heat. If \( dQ \) is larger than \( dQ^E \), the temperature gradually increases. A high temperature means that the average kinetic energy of a clique is high, indicating that the variation in colors of the corresponding pixel is also high. At such a pixel, although the difference between \( b_{c,i,j,t} \) and \( X_{c,i,j,t} \) is large, the effect, \( dS_{i,j,t} \), is abated due to the higher temperature. Thus, repeated motions such as waving leaves are correctly classified as a part of the background.

Motion segmentation using CNF adapts to the elapse of time or illumination changes and can appropriately treat repeated motions. Moreover, it is robust to noise produced by
null
3.3 Defining $T'_k(Q)$

$T'_k,ij_t$ in (6) is an absolute temperature and cannot take a negative value [21]. Let $\alpha$ be the absolute temperature of the surrounding atmosphere. $T'_k,ij_t$ cannot be smaller than $\alpha$ because it is in thermal equilibrium when $T'_k,ij_t = \alpha$. Thus, $T'_k,ij_t(Q)$ is defined as follows in LCNF:

$$T'_k,ij_t = (1 - \rho)T'_k,ij_{t-1} + \frac{dQ'_k,ij_t}{n_k,ij_t}, \quad T'_k,ij_t > \alpha. \quad (9)$$

In a system without $\alpha$, the temperatures may decrease near zero. This means that $dS$ can take an extremely high value even though $dQ$ is small. The effect of employing $\alpha$ is described in the experiment section.

3.4 Automatic Normalization

If $\alpha$ is chosen as a value greater than 1, $dS''$ of LCNF is normalized in a specific range even though there are no additional normalization steps for the following reasons:

- the total number of molecules that comprise a 2nd-order clique (pixel) is chosen as $K \times n_i$;
- the sum of weights of channels, $w_c$, is constant; and
- $(X_{c,ij_t} - b_{k,c,ij_t})^2$ is within the range of $0 - 255^2$.

Note that the sum of normalized values is also a normalized value. Thus, $p'i(i,j,t)$ in the paper [19] is always automatically normalized.

4. Experiment and Analysis

The system requires only 48 ms per frame for a sequence of 720p (1280 \times 720) color images on a computer equipped with a 2.8 GHz CPU and 685 MHz graphics card. We used a comparatively heavy algorithm, but the algorithm can be run in parallel. It can be executed in essentially one-half the time with two processors running simultaneously. We utilize a graphics processing unit (GPU) for the computation, and the system runs at 21 frames per second. This is faster than other recent techniques, as shown in Table 1. Although the other approaches generally use a resolution under 352 \times 240, our method (LCNF) is much faster.

4.1 Experimental Datasets

We installed cameras indoors and outdoors to confirm adaptability to waving trees, monitor flicker, and slowly moving shadows. We verified that the system clearly segments moving objects from backgrounds and effectively eliminates noise under situations where the cameras are slightly shaken by the wind or human movement. For these experiments, we took four video clips using stationary or moving cameras. The first video clip comprises 10189 frames containing four cars and nine pedestrian groups taken by a stationary camera, while the second video clip comprises 20121 frames containing twenty-eight cars and seventeen pedestrian groups taken by a stationary camera. Representative frames of these clips are depicted in Figs. 3 (a) and 3 (b). The third video clip comprises 473 frames containing only one subject taken by a stationary camera. The colors and brightness levels of the backgrounds are similar to the upper garment of the subject, as shown in Fig. 3 (c). The last video clip comprises 1001 frames containing a moving blue pen that has a desktop as a background, and was taken using a moving camera. The camera starts to move from the 211st frame. The background desktop contains pens and a monitor that have colors and brightness levels similar to those of the tracked pen, as shown in Fig. 3 (d).

4.2 Effect of Adopting $w_c$

As described in the paper [19], we also adopt both the spatial function, $N(i,j)$, and the temporal function, $G(t)$. For the first experiment, we use the Kronecker delta function at both $N(i,j)$ and $G(t)$ for the images in Figs. 4 and 5. Thus, the system considers neither spatial nor temporal information. Figure 4 shows the results of a system using simple CNF, and illustrates the effects of $w_c$ noted in Sect. 3.
uppermost image is an input image and the middle image is a result of the algorithm developed by Fazli et al. [16]. We can see that some parts of the body are misclassified as part of the background. The original background color of the misclassified region is white, and the color of the upper garment is light yellow. The colors are distinctly different from each other but the brightness levels (intensities in gray scale) are similar. Whereas other algorithms, including those in standard GMM [1], GMM with MRF (GnM) [12], improved GMM (IGMM) [16], used identical weights for all color channels, we reduce the weight of the brightness channel and increase the weight of color channels in YCbCr. Since moving objects usually differ more in color than in brightness relative to the background, the use of appropriate weights \( w_c \) leads to more accurate results, as illustrated in the bottommost image in Fig. 4. Henceforth, all results of our method employ \( w_c \), as used in the images shown in Fig. 4.

4.3 Effect of Considering \( \alpha \)

Figure 5 shows the results considering \( \alpha \). The uppermost image is an input image taken when the camera was slightly shaken by the wind. Only the objects inside the red circle are moving objects and all others should be classified as part of the background. The middle image is the result of a system that does not use \( \alpha \), while a system using the algorithm introduced in IGMM [16] also generates results similar to the middle image, which contains a great deal of noise. The bottom image is the result of using one as a \( \alpha \) value, and we can see that the noise is substantially removed. Henceforth, all results of our method employ the same \( \alpha \), as used in the images shown in Fig. 5.

4.4 Considering Spatial Information

The input images presented in Fig. 6 are also captured from a camera that was slightly shaken by the user. The three left images are the results of the algorithm from GnM [12]. They employ MRF as a post-processing step in addition to the GMM method. Some noise is eliminated using MRF,
Fig. 7 Resultant images of LCNF. The results consider both spatial and temporal information: (a) input image, (b) zoomed-in image, (c) resultant image of GMM [1], and (d) resultant image of LCNF. The white regions in (d) are classified as moving objects when the system considers spatial information only, and the regions including both white and grey regions are classified as moving objects when the system considers both spatial and temporal information. (e) and (f) are also resultant images of GMM [1] and LCNF, respectively.

Fig. 8 Resultant images of LCNF. Motion segmentation results of bottommost images are better than the second row images.

Fig. 9 Resultant images of LCNF. We apply red fluorescence effects to moving objects, which are indicated by red circles.

but the images still contain substantial noise. We use the Kronecker delta function only at $G(t)$, which means that the system considers only spatial information. The system eliminates most noise by using $w_c$ and $\alpha$, and segments a moving object comparatively clearly. Moreover, our method is faster than the algorithm used in GnM [12]. We implement the algorithm of GnM [12] through the use of a GPU, but the fps is less than that of our method, as shown in Table 1.
4.5 Considering both Spatial and Temporal Information

We consider both spatial and temporal information using LCNF, the results of which are shown in Figs. 7 and 8. Figure 7 (a) is an input image, and Fig. 7 (b) shows a zoomed-in image of the truck in Fig. 7 (a). Figure 7 (c) is the result of when the algorithm from GMM [1] is applied to the truck, and Fig. 7 (d) is the result of using LCNF. The white regions in Fig. 7 (d) are classified as moving objects when the system considers spatial information only, and the regions including both white and grey regions are classified as moving objects when the system considers both spatial and temporal information. Figures 7 (e) and 7 (f) also show the same effects of LCNF. These images are the results of segmenting the pedestrians, and the right image is the result of using LCNF.

The two uppermost images in Fig. 8 are input images. There are some waving trees and gradually moving shadows in the left image. In the right image, the camera is slightly shaken by the user. The images in the second row show noise in the trees, cars, and calendar. The bottommost images are the results of using LCNF, in which the system clearly segments only the moving objects. The system does not miss any moving objects and successfully eliminates all noise.

4.6 Tracking Experiments

To show the qualitative results of the proposed method (LCNF), we apply fluorescence effects to the sampled images, as shown in Figs. 9 and 10. These images show that the proposed method clearly separates meaningful moving objects from the backgrounds, including waving trees and slowly moving shadows, and effectively eliminates noise equivalent to a level achieved through the application of refined post-processing steps on the results of traditional motion segmentations. The parameters used for these results...
Table 2  Comparison of the tracking success rates.

|                  | GMM [11] | GnM [12] | IGMM [16] | DC [29] | CNF  | LCNF  |
|------------------|----------|----------|-----------|---------|------|-------|
| **First Video Clip** |          |          |           |         |      |       |
| MS               | 30.77 %  | 76.92 %  | 76.92 %   | 83.12 % | 82.88 % | 84.62 % |
| CS               | 38.46 %  | 84.62 %  | 76.92 %   | 90.56 % | 91.52 % | 92.30 % |
| AS               | 30.77 %  | 84.62 %  | 84.62 %   | 85.19 % | 84.62 % | 84.62 % |
| **Second Video Clip** |        |          |           |         |      |       |
| MS               | 57.78 %  | 82.22 %  | 80.00 %   | 88.31 % | 91.11 % | 91.11 % |
| CS               | 68.89 %  | 88.89 %  | 84.44 %   | 88.89 % | 84.44 % | 88.89 % |
| AS               | 73.33 %  | 75.56 %  | 93.33 %   | 94.43 % | 95.56 % |

Fig. 11  A white region which has to be modeled for object tracking. (a) Input images. (b) Resultant images of motion segmentation.

are as follows:

\[ K \to 4, \quad \lambda \to 0.01, \quad \rho \to 0.007, \quad \kappa \to 0.005, \quad \alpha \to 1 \quad (10) \]

To show the quantitative results, we tested the resulting images using MeanShift (MS) [26], CamShift (CS) [27], and ABCShift (AS) [28] object trackers since we do not have ground truth data. A tracked object is modeled as having a class conditional color distribution, \( P(\bar{C}|O) \), for each pixel with color \( \bar{C} \), which is the probability of the color of the pixel, given that the pixel belongs to the tracked object, \( O \). This object distribution is usually learned offline from training images, and the shapes of the regions of learned objects are geometrical, such as rectangles or circles. These geometrical regions are certain to contain some background regions together. However, we model the object distribution using the results of motion segmentation as shown in Fig. 11. Figure 11 (a) contains input images and Fig. 11 (b) contains segmented results. The white regions in the yellow rectangles of Fig. 11 (b) are the foreground regions that have to be modeled. The regions have a greater chance to contain only regions of moving objects. Thus, it could be seen that the tracking performances or success rates reach the accuracies of the motion segmentation results. The motion segmentation algorithms used for the quantitative experiments are GMM [11], GnM [12], IGMM [16], DECOLOR (DC) [29], and the proposed method.

Table 2 shows the tracking success rates according to each motion segmentation and tracking algorithm for the first and second video clips. The proposed method shows generally superior results over the other methods. Table 3 shows the frame number when the object trackers (MeanShift, CamShift, and ABCShift) begin to miss the subject in the third video clip. We start to track the subject from the 146th frame. Table 4 shows the frame number when the object trackers begin to miss the pen in the last video clip. We start to track the pen from the 50th frame.

5. Conclusions

We have shown a novel and entropy-based probabilistic method for motion segmentation. It includes designing a CNF model for utilizing spatio-temporal image information. In [19], CNF was originally proposed for solving uncalibrated stereo matching problems. However, we designed the CNF model to be suitable for motion segmentation. In addition, we proposed an advanced model, Layered-CNF, to make a more suitable CNF model for a real-time motion segmentation method that is stable and robust.

This method deals with repeated background movement and slow shadow movements occurring by lighting changes. It can treat multimodal distributions caused by monitor flickers, waving branches, shadows, and other troublesome features of the actual world often mentioned as difficult problems in computer vision.
A system utilizing CNF clearly separates moving objects from the backgrounds and effectively eliminates noise, equivalent to a level achieved through the application of refined post-processing steps on the results of motion segmentation. Our system has also been successfully used to track people in indoor environments, as well as people or cars in outdoor environments.

There are several possible directions for future research. We are interested in applying CNF to other time-varying imagery and designing well-defined models that optimally fit the applications, such as LCNF, for motion segmentation. Developing algorithms that improve the performance of our method will also be another area of future work.

References

[1] C. Stauffer and W.E.L. Grimson, “Learning patterns of activity using real-time tracking,” IEEE Trans. Pattern Anal. Mach. Intell., vol.22, no.8, pp.747–758, 2000.
[2] A. Cavallo, O. Steiger, and T. Ebrahimi, “Semantic video analysis for adaptive content delivery and automatic description,” IEEE Trans. Circuits Syst. Video Technol., vol.15, no.10, pp.1200–1209, 2005.
[3] T. Moeslund, A. Hilton, and V. Kruger, “A survey of advances in vision-based human motion capture and analysis,” Comput. Vis. Image Understand., vol.104, no.2, pp.90–126, 2006.
[4] Y. Lim, K. Park, J. Lee, S. Aoki, and F. Fernando, “MMT: An emerging MPEG standard for multimedia delivery over the internet,” IEEE Multimedia, vol.20, no.1, pp.80–85, 2013.
[5] D. Forsyth and J. Ponce, Computer Vision: A Modern Approach, Englewood Cliffs, Prentice-Hall, NJ, 2003.
[6] B. Horn and B. Schunck, “Determining optical flow,” Artif. Intell., vol.17, pp.185–203, 1981.
[7] J. Barron, D. Fleet, and S. Beauchemin, “Performance of optical flow techniques,” IJCV, vol.12, no.1, pp.43–77, 1994.
[8] F. Lian, Y. Lin, C. Kuo, and J. Jean, “Voting-based motion estimation for real-time video transmission in networked mobile camera systems,” IEEE Trans. Industrial Informatics, vol.9, no.1, pp.172–180, 2013.
[9] Y. Weiss, “Smoothness in layers: Motion segmentation estimation using NonParametric mixture,” CVPR, pp.520–526, 1997.
[10] A.S. Ogale, C. Fermuller, and Y. Aloimonos, “Motion segmentation using occlusions,” IEEE Trans. Pattern Anal. Mach. Intell., vol.27, no.6, pp.988–992, 2005.
[11] J.D. Joseph, S. Jayathu, and R. Ranga, “Appearance based tracking with background subtraction,” ICCSE, pp.643–659, 2013.
[12] J. Migdal and W.E.L. Grimson, “Background subtraction using Markov thresholds,” WACV/MOTION, pp.58–65, 2005.
[13] T. Kentaro, K. John, B. Barry, et al., “Wallflower: Principles and practice of background maintenance,” ICCV, vol.1, pp.255–261, 1999.
[14] S. Geman and D. Geman, “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images,” IEEE Trans. Pattern Anal. Mach. Intell., vol.6, no.6, pp.721–741, 1984.
[15] S. Della Pietra, V. Della Pietra, and J. Laﬀerty, “Inducing features of random fields,” IEEE Trans. Pattern Anal. Mach. Intell., vol.19, no.4, pp.380–393, 1997.
[16] S. Faizi, H.M. Pour, and H. Bouzari, “Multiple object tracking using improved GMM-based motion segmentation,” ECTI-CON 09, pp.1130–1133, 2009.
[17] S.G. Brush, Lectures on Gas Theory, New York, 1995.
[18] J. Daintith, Oxford Dictionary of Physics, Oxford University Press, 2005.

Eunjin Koh received his BS and MS in computer science from Inha University, South Korea, in 2005 and 2007, respectively and received his PhD from University of Science and Technology, Daejeon, South Korea, in 2011. Since 2011, he has been a member of Agency for Defense Development, Daejeon, South Korea. His main research interests include image processing, human-computer interaction, object tracking, and pattern classification.

Chanyoung Lee received his B.Sc. in Mathematical Education from Sungkyunkwan University, Seoul, Republic of Korea in 2005. In 2007 and 2011, he received his M.Sc. and Ph.D. in Applied Mathematics from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, respectively. His main research interests include image processing, computer vision, object tracking, and pattern classification.
Dong Gil Jeong received his BS in electricity and electronic engineering from Pusan National University, Busan, South Korea, in 2003 and received his MS from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2005. Since 2005, he has been a member of Agency for Defense Development, Daejeon, South Korea. His main research interests include image processing, object tracking, synthetic image generation and integrated simulation.