コンピュータビジョンと
イメージメディア 101－2
（1996．9．12）

確率的モデリングに基づくRGB動画像の局所動的領域分割

クリアンスキ アダム　　長橋　宏
東京工業大学 大学院総合理工学研究科

本論文では，RGB動画像の動的領域分割手法について述べる．局所動的情報，即ち，画素の時間的，空間的隣接部分の動きの知識の確率的モデリングに基づいた動物体の双峰確率表現手法を提案する．実験により，確率モデルを双峰表現にしたところより良好な動的領域分割の結果が得られた．本論文では，モデリングの簡単化，即ち，比較的簡単な分布関数の任意のRGB動画像への応用のため，MRF最適化手法を使用している．

Local motion segmentation in RGB image sequences based on stochastic modeling

Adam KURIAŃSKI　Hiroshi NAGAHASHI
Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology

Abstract

A problem of motion segmentation in RGB image sequences is addressed. Methods considered are based on a stochastic modeling of a local motion information, i.e. the knowledge about motion available in a pixel's neighbourhood in space and time domains. A bimodal approach to the stochastic description of moving objects is presented. It is shown by examples that the increase of mode number of the stochastic model results in motion segmentation of better quality. A MRF optimization framework is used in order to tackle a modeling simplification, i.e. the application of few relatively simple distribution functions to any RGB motion sequence.

1 Introduction

Motion analysis plays an important role in computer vision due to the relevance to contemporary applications. Image sequence coding, monitoring, target tracking are examples of topics which are under interest of researchers nowadays. Even a method originated in motion analysis has been used in human face recognition[Bey96].

In motion analysis, as in almost all applications of computer vision, two major algorithm groups exist: local algorithms and global ones. The latter use the results found by the former which operate on pure imaging data, i.e. the values of brightness information in a pixel. The aim of this paper is to show a method of localization of moving object in a sequence of RGB frames based on stochastic approach and local modeling only. A historical profile of the method will be given; from the original algorithm of Bouthemy and Lalande for gray-level frames, to the recent suggestion based on a bimodal model used for motion segmentation in RGB sequences.

The notion *"local motion analysis"* is understood as no need of any a priori information of shape, size, velocity, and direction of the moving object. Only a brightness information as well as a difference of brightness in the time domain are used. The knowledge of the phenomenon of motion, i.e. how to distinguish the object from the background, is expressed in a model construction. Model parameters are responsible for matching the algorithm to a sequence analyzed.

Generally, the methods of local motion analysis can be divided into two groups: based on optical flow and based on "pure" pixel labeling. The approach addressed in this paper belongs to the second group. On one hand, avoiding the detection of optical flow eliminates one step from the processing algorithm, on the other hand, the information about the direction of motion is lost. One should notice that resigning from optical flow does not mean that all problems related to its detection have disappeared, for example *the aperture problem*[Hor91] also exists in labeling approach and is called *an ambiguity of motion information.* To tackle this problem in our methods an extended local motion information vector is used which in the case of color RGB sequences consists of six components.

In following sections first, the methods based on unimodal models are presented, from the Lalande–Bouthemy's proposition to the contemporary algorithms. Next, a bimodal model extension is described. Both types of algorithms are discussed and advantages involved by model extension are noticed. Some experimental results are given as well.

2 Unimodal approaches

Every local motion analysis method consists in the application of at least two consecutive sequence frames. Such methods originated from a simple thresholding of the difference of brightness of two frames. The result sometimes called a *change mask* relates both consecutive frames analyzed. There exist also more sophisticated approaches to the

detection of change masks, for example presented in[HNR84,SJ89] which use also the values of neighbours of a pixel.

Every change detection method results in a limited motion information and the localization of moving object in each frame is usually hardly possible. Next step was done by Bouthemy and Lalande[BL90] who suggested a post–processing method using the change masks as an input information and a statistical unimodal description of the difference of brightness to localize moving objects in every frame of a gray-level image sequence. The method was accompanied by Markov random field framework to achieve a connectivity of the mask of moving object detected. The Lalande–Bouthemy model has involved a series of models. From local motion modeling point of view, a very important Lalande and Bouthemy's suggestion was to model separately three subareas resulted from the object motion, i.e. *covered background*, *recovered background*, and *overlapping situation*. They however, used the same statistical distribution in both transient situations (covered and recovered background) which really limited the applicability. A separate modeling of both transient situations was introduced in paper.[Kur94] Moreover, the experiments performed and presented in[Kur94] allowed one to introduce a new modeling approach to local motion labeling, i.e. *a dependent use of brightness difference and brightness*[KN95]. This modification involved a significant improvement of the quality of motion segmentation results. Even motion areas not found by a change detection method were recovered by the segmentation algorithm.

The first autonomous motion segmentation method, i.e. without the use of change masks, was introduced in[KAN95] and the first RGB version of it was presented in.[KAN96a]

All the models, what was not written clearly in mentioned papers, were in fact unimodal ones, i.e. based on an assumption that the brightness information of this part of image which shows moving object is distributed according to a Gaussian function. If so, the difference of brightness is also Gaussian and moreover, the total distribution is Gaussian as well. According to an intuitive knowledge, the description of brightness of moving object by one Gaussian function may not be an exact model in any case. This is the reason why a MRF framework has to be incorporated to the model, i.e. in order to "correct" errors of labeling resulted from too simple model. In fact, both unimodal modeling and MRF framework together involved a tradeoff between the model complexity and the quality of segmentation.

In short, the method of motion segmentation in RGB sequences based on the unimodal model looks as follows[1]:

- An RGB sequence is considered as a sequence of pairs of consecutive frames, i.e. a sequence of two-frame subsequences. Moreover, there is an overlapping of subsequences, i.e. one frame of source sequence belongs to two neighbouring subsequences.
- A local motion information vector consists of six components:

$$\boldsymbol{o}_k = [\Delta R_k, \Delta G_k, \Delta B_k, R_k, G_k, B_k]^T \tag{1}$$

[1]for full description see paper[KAN96b]

where, k is a number of sequence frame (the index of subsequence), i points a pixel in the image plane, $\Delta R_k = R_{k+1} - R_k$ means the difference of R component of k and $k+1$ frames, superscript T is a symbol of vector transposition.

- A stochastic model of one subsequence consists of four Gaussian distributions functions describing the local motion information $\boldsymbol{o}_k$. The number of Gaussian distributions is related to four types of motion areas which exist in two consecutive frames: *covered background, recovered background, overlapping situation* and *static background*. In a simplest version (without MRF framework) if there are parameters of all four distributions, in every pixel a label configuration is chosen which results in the largest likelihood, i.e. this is a Maximum Likelihood (ML) labeling approach.
- A MRF framework is used which supports the spatial connectivity of a label set detected. The framework removes misdetections caused by the model simplification (unimodal modeling) as well as noise and so on. Exponential expressions of four mentioned Gaussian distributions are used as additional energy term of MRF. A deterministic relaxation algorithm is applied to find a mode of MRF model, i.e. to perform a motion segmentation.

From mathematical point of view, the model and the MRF framework applied to two consecutive frames look as follows:

- distribution functions of $\boldsymbol{o}_k$ vector in two-frame subsequence:

$$p(\boldsymbol{o}_k \mid l) = \frac{1}{(2\pi)^3} \exp\left\{-\frac{1}{2}\left[(\boldsymbol{o}_k - \boldsymbol{\mu}_l)^T \boldsymbol{\Sigma}_l^{-1} (\boldsymbol{o}_k - \boldsymbol{\mu}_l)\right] - \frac{1}{2}\ln(\mid \boldsymbol{\Sigma}_l \mid)\right\} \tag{2}$$

where, $l \in \{0,1,2,3\}$ is an index which points a label configuration, i.e. *(background, background)*, *(motion, background)*, *(background, motion)*, and *(motion, motion)*, respectively,

$$\boldsymbol{\mu} = [\overline{\Delta R}, \overline{\Delta G}, \overline{\Delta B}, \overline{R}, \overline{G}, \overline{B}] \tag{3}$$

denotes a mean vector, $\boldsymbol{\Sigma}$ is a 6×6 covariance matrix of the vector $\boldsymbol{o}_k$. If Eqs.(2) are considered as a function of l parameter, they become a likelihood function.

- global energy of MRF:

$$W = W_{s_k} + W_{s_{k+1}} + W_{c_k} \tag{4}$$

where, W_{s_k}, $W_{s_{k+1}}$ are the spatial energies responsible for the connectivity of configuration of labels of two consecutive frames k and $k+1$, respectively, and W_{c_k} is a consistency energy which describes the influence of vector $\boldsymbol{o}_k$.

- local energy of MRF:

$$U(i) = U_{s_k}(i) + U_{s_{k+1}}(i) + U_{c_k}(i) \tag{5}$$

where, all local sub-energies $U_{s_k}(i)$, $U_{s_{k+1}}(i)$, $U_{c_k}(i)$ are only the parts of energies W_{s_k}, $W_{s_{k+1}}$, W_{c_k} that include the potentials of all cliques to whom pixel i belongs. Both spatial energies assume the form:

$$U_s(i) = \sum_{cs \in C} V_{cs} \tag{6}$$

where, C is the set of all cliques including pixel i, cs denotes a two-pixel spatial clique, V_{cs} denotes the potential of cs which equals:

$$V_{cs} = \begin{cases} \beta_s & \textit{if both labels are different,} \\ -\beta_s & \textit{if both labels are the same,} \end{cases} \tag{7}$$

where β_s is a constant grater than 0. The consistency subenergy is:

$$U_{c_k}[l(i)] = \frac{1}{2}\left[\boldsymbol{o}_k(i) - \boldsymbol{\mu}_l\right]^T \boldsymbol{\Sigma}_l^{-1}\left[\boldsymbol{o}_k(i) - \boldsymbol{\mu}_l\right] + \frac{1}{2}\ln\left(\mid \boldsymbol{\Sigma}_l \mid\right) \tag{8}$$

where, l is an index of label configuration.

3 Bimodal approach

In order to enhance the fitting quality of the mathematical model to the data analyzed we have suggested[KN96] a model featured with two modes of brightness information of moving object, i.e. we have assumed a priori that the moving object consists of bright and dark areas. The bright areas are described by the first mode and dark areas by the second one.

Virtually, it would be possible to use a bimodal distribution as a model of brightness of moving object however, in such a case one usually would be able to tell nothing about neither the shape and the type of the distribution of brightness difference nor the type of the total distribution of brightness and brightness difference. Moreover, if the total distribution does not belong to the exponential family, a MRF optimization framework would not be applicable[Li95]. So, we approached the problem in other way. We assumed that the bright areas of the object are modeled by one Gaussian distribution and the dark areas by another Gaussian with different parameters of course. Due to the last assumption the difference of brightness is also Gaussian as well as the total distribution is.

In other words, the basic modeling assumption in a new approach is that one RGB frame can be described in total by three Gaussian distributions; one used for background, and two others for moving object. This involves three-label set used by motion segmentation algorithm applied to a two-frame subsequence. The other modeling and motion segmentation algorithm assumptions are:

- vector $\boldsymbol{o}_k$ given by Eq.(1) is used as the local motion information,
- as a consequence of three-label set, there are nine possible label configurations in a pixel:

  (background, background), (background, motion1), (background, motion2),
  (motion1, background), (motion1, motion1), (motion1, motion2),
  (motion2, background), (motion2, motion1), (motion2, motion2)

  where the first label is assigned to a pixel in the first frame of a two-frame subsequence and the second label to the same pixel but in the consecutive frame.

- the model of a subsequence of two consecutive frames consists of nine distributions given by Eq.2 which involves $l \in \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$ where l in the index of label configurations given above.

In the model presented, there are four label configurations responsible for the transitive motion situations (all the label configuration where exactly one *background* label exists), four label configurations corresponding to the overlapping situation (no *background* label), and one where the static background exists in both frames.

The other mathematical foundations remain the same as for model presented in Sec.2 with an extension of the possible values of index l where applicable.

4 A comparison of unimodal and bimodal approaches by experiments

We performed a series of experiments with different RGB image sequences. An example of test sequence is shown in Fig.1. The moving object consists of two different types of color areas, white car body and dark windows, shadow and wheels. For the test sequence we applied two motion segmentation algorithms, one based on the description from Sec.2 and the other presented in Sec.3. The results of experiments are shown in Figs.2, and 3, respectively. In the case of the unimodal approach the moving shadow, wheels, car windows, and car body were modeled by one Gaussian distribution, whereas in the bimodal algorithm for the car body and the dark moving areas different modes were applied.

The results found by the bimodal approach are of better quality. Due to an initialization step[KAN96b] the first mask found by both algorithms are of worse quality however, in the case of the bimodal model the exact shape of the object detected is achieved earlier than in the case of the unimodal approach. The fourth frame found by the bimodal algorithm shows an exact shape of the object whereas the frame of the same number detected by the unimodal model still has an undetected areas of moving shadow. Moreover, in all frames found by the unimodal algorithm there exists the undetected area of the right lamp of moving car. Such an undetected area disappeared in the bimodal algorithm already in the third frame. The undetected lamp area deteriorates the connectivity of the mask of moving object.

As a conclusion one can notice that the motion segmentation algorithm based on the bimodal model results in the better quality masks of moving object and the smaller influence of the initialization step.

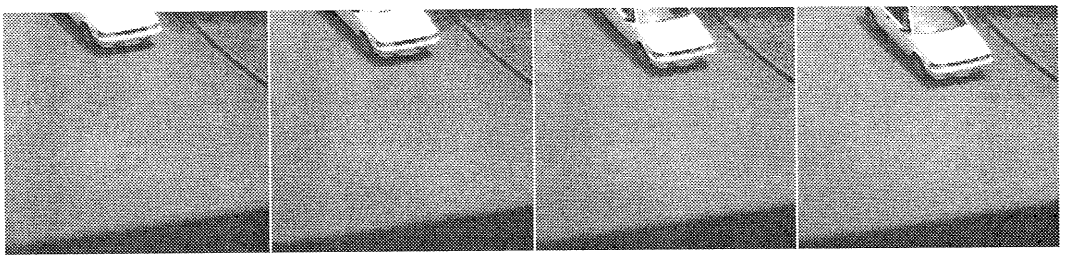

Figure 1: RGB test sequence.

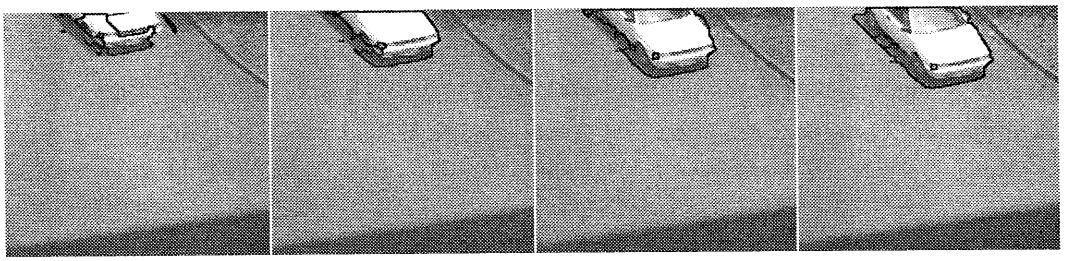

Figure 2: Motion segmentation with unimodal modeling.

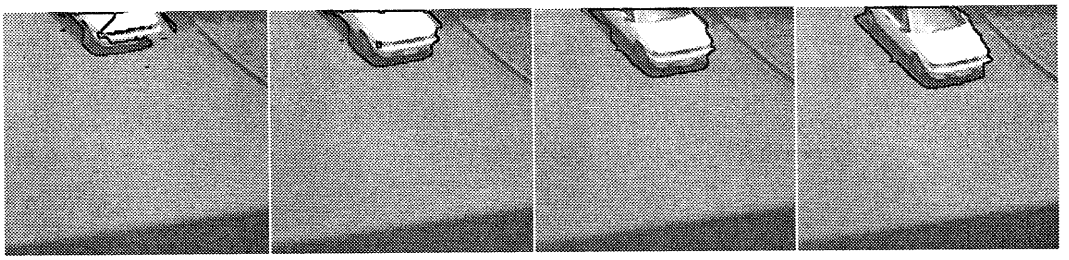

Figure 3: Motion segmentation with bimodal modeling.

5 Conclusions

Two kinds of stochastic local motion segmentation algorithms by labeling approach are presented in this paper. One method is based on a unimodal model and the other one on a bimodal description. Performed experiments have shown that the bimodal approach is better than the unimodal one because the influence of the initialization step is shorter and masks detected are of better quality.

6 REFERENCES

[Bey96] D. Beymer. Pose invariant face recognition using real and virtual views. Technical Report AI 1574, Massachusetts Institute of Technology, MIT, March 1996. PhD thesis.

[BL90] P. Bouthemy and P. Lalande. Detection and tracking of moving objects based on a statistical regularization method in space and time. In *Proceedings of European Conference on Computer Vision*, pages 307–311, Antibes, France, April 1990.

[HNR84] Y. Z. Hsu, H. H. Nagel, and G. Rekers. New likelihood test method for change detection in image sequence. *Computer Vision Graphics and Image Processing*, 26:73–106, 1984.

[Hor91] Berthold K. P. Horn. *Robot Vision.* The MIT Press, 7th edition, 1991.

[KAN95] A. Kuriański, T. Agui, and H. Nagahashi. Turning face localization based on hidden MRF model with two sources of observation and 2D Gaussian distribution. In *Second Asian Conference on Computer Vision ACCV'95*, volume 3, pages 422–426, Singapore, December 5–8 1995.

[KAN96a] A. Kuriański, T. Agui, and H. Nagahashi. Motion segmentation in RGB image sequence based on hidden MRF and 6D Gaussian distribution. In *Proceedings of VCIP'96*, pages 657–667, Orlando, USA, March 17–20 1996. SPIE.

[KAN96b] A. Kuriański, T. Agui, and H. Nagahashi. Motion segmentation in RGB image sequence based on stochastic modeling. *IEICE Transactions on Information and Systems*, 1996. Submitted for publication, and revised.

[KN95] A. Kuriański and M. Nieniewski. A model of the MRF with three observation sources for obtaining the masks of moving objects. In Gunilla Borgefors, editor, *9th Scandinavian Conference on Image Analysis*, pages 931–940, Uppsala, Sweden, June 6–9, vol.2 1995. IAPR.

[KN96] A. Kuriański and H. Nagahashi. Detection of moving objects based on bimodal stochastic modeling. In *Proceedings of the Spring 1996 Information and Systems Society Conference of IEICE*, Kanazawa, Japan, September 18–21 1996. IEICE.

[Kur94] A. Kuriański. Detection and motion tracking by means of spatio-temporal modeling of images using random fields. Technical Report 37, Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland, December 1994. PhD thesis in Polish.

[Li95] S. Z. Li. *Markov Random Field Modeling in Computer Vision.* Computer Science Workbench. Springer-Verlag, Tokyo, 1st edition, 1995.

[SJ89] K. Skifstad and R. Jain. Illumination independent change detection for real world image sequences. *Computer Vision Graphics and Image Processing*, 46:387–399, 1989.