Image and Vision Computing 28 (2010) 376–390

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Colour, texture, and motion in level set based segmentation and tracking

Thomas Brox [a,*], Mikaël Rousson [b], Rachid Deriche [c], Joachim Weickert [d]

[a] University of California at Berkeley, Surtardja Dai Hall, Berkeley, CA 94720-1758, USA
[b] Polar Rose, Romix, BP 65, 06902 Sophia Antipolis Cedex, France
[c] Odyssée Project Team – INRIA/ENPC/ENS, INRIA, 2004 Route des Lucioles, BP 93, 06902 Sophia Antipolis, France
[d] Mathematical Image Analysis Group, Building E11, Saarland University, 66041 Saarbrücken, Germany

ARTICLE INFO

Article history:
Received 29 July 2005
Received in revised form 27 June 2007
Accepted 6 June 2009

Keywords:
Image segmentation
Tracking
Level set methods
Nonlinear diffusion
Texture
Motion

ABSTRACT

This paper introduces an approach for the extraction and combination of different cues in a level set based image segmentation framework. Apart from the image grey value or colour, we suggest to add its spatial and temporal variations, which may provide important further characteristics. It often turns out that the combination of colour, texture, and motion permits to distinguish object regions that cannot be separated by one cue alone. We propose a two-step approach. In the first stage, the input features are extracted and enhanced by applying coupled nonlinear diffusion. This ensures coherence between the channels and deals with outliers. We use a nonlinear diffusion technique, closely related to total variation flow, but being strictly edge enhancing. The resulting features are then employed for a vector-valued front propagation based on level sets and statistical region models that approximate the distributions of each feature. The application of this approach to two-phase segmentation is followed by an extension to the tracking of multiple objects in image sequences.

1. Introduction

The segmentation and tracking of objects in image sequences is a challenging problem in image analysis. Robust algorithms exist for many industrial applications, yet efficiency and robustness are most of the time obtained by adding strong constraints issued from prior knowledge and the controlling of the environmental conditions. Numerous approaches for tracking, for instance, propose to use models of the moving objects or of the background. Although such approaches usually turn out to be quite robust, they often lack generality. Relying on sharp constraints, they are designed to tackle precisely defined problems. Moreover, they usually need a supervised learning stage, and their performance highly depends on the quality of the learning samples.

In this paper, we want to use as little prior knowledge as possible. In order to succeed in the task of image segmentation, it is therefore necessary to use as much information of an image as possible. Thereby, the combination of different cues is essential for the robustness of segmentation and tracking. The combination is to be designed such that if one cue is corrupted, the method can still rely on the remaining information.

Before one can combine different types of information, this information has to be extracted first. There is no such problem in case of primary features like the grey value or colour, which are already given as input data. However, as soon as secondary features like texture or motion play a role, there are many possibilities how to extract them from the image. In the ideal case, the features should be highly discriminative while inducing as little data load as possible.

In the field of texture analysis the most frequently used methods are based on Gabor filters [31,47,58,59]. Neurobiology indicates Gabor filters to be important in animal and human vision. However, they have the decisive drawback to induce a lot of redundancy and thus many feature channels. Similar problems appear with texture features based on Markov Random Fields (MRFs) [24]. As soon as a MRF of reasonable order is used, there arise many parameters not only causing lots of feature channels but also problems in estimating them. An alterative has been proposed in [6] where the structure tensor, also known as second moment matrix, is used for discriminating textures. It yields only three feature channels and thus provides significantly larger information content per channel than Gabor features. Recently, the structure tensor has been extended to so-called nonlinear structure tensors based on nonlinear diffusion [11,14], which are data-adaptive and therefore more accurate in the vicinity of discontinuities. In the present paper, we make use of this concept in order to extract texture features and to combine them with other cues.

Considering motion, the optical flow is the principal method to extract this information. Optical flow estimation is a complete research area on its own, and there exist plenty of different techniques; see e.g. [4,42,62] for overviews. Interestingly, the

* Corresponding author.
E-mail addresses: brox@eecs.berkeley.edu (T. Brox), mikael.rousson@gmail.com (M. Rousson), Rachid.Deriche@sophia.inria.fr (R. Deriche), weickert@mia.uni-saarland.de (J. Weickert).

0262-8856/$ - see front matter
doi:10.1016/j.imavis.2009.06.009

nonlinear structure tensor, already applied for deriving texture features, can also be used here [14].

The combination of features is realised at two different stages (see Fig. 1): first, nonlinear diffusion is applied to the extracted feature vector, thus inducing a coupled smoothing of all feature channels. This permits to deal with possible outliers, and the spatial coherence between various information is highly improved. Afterwards, these smoothed versions of the features are combined in joint probability density functions that describe the interior of regions. These densities are approximated by Gaussian densities and nonparametric Parzen estimates [37,50]. We will show that by choosing appropriate statistical models one obtains an implicit weighting of the features by their discriminative power.

Maximising the total *a-posteriori* probability finally leads to a partitioning of the image domain. For this optimisation problem, we employ level set methods [26,44,45,60]. The earlier level set works are on edge based active contour models [16,17,36,40]. They have later been extended to region based models [18,19,47,49], which can be related to the energy functionals proposed by Mumford and Shah [43] and Zhu and Yuille [73]. A recent survey on level set based segmentation can be found in [23]. Formulating these functionals in the level set framework has several advantages. First of all, the embedding of the 1-D curve into the 2-D level set function as its zero-level line allows for combining curve constraints with region constraints. Furthermore, topological changes can easily be handled with level set representations. In particular, regions need not necessarily be connected and parts of a region can split or merge. Finally, the optimisation of the partitioning problem by evolving curves is perfectly suited for applying the method to tracking problems: once the object contour in one image is found, only few iterations are necessary to adapt the curve to the shape and position of the object in the next frame.

1.1. Related work

The integration of different types of information has been extensively studied in various fields in the past. Here, image cues will stand for the information to integrate, but it may also be the fusion of data obtained from different sensors, which is a whole domain in itself.

One possibility is to create a feature vector from all the cues and run classical clustering algorithms in the resulting feature space [5,46]. In [5], a feature space for image segmentation is obtained from the extraction of texture and colour cues. Then, an EM algorithm is employed to estimate the parameters of a Gaussian mixture in the corresponding 6-D feature space. Similarly, it has been suggested to maximise the *a-posteriori* probability, or equivalently, to minimise the negative log-likelihood [48,73]. With the assumption of no correlation between cues, a global energy is defined by summing the log-likelihoods for each single cue. Auto-adjustment of these weights has been proposed in the case of image segmentation in [33,35] and for visual tracking in [61]. Martin et al. [41] propose a boundary detector based on different cues, where the weights are learned from a human-labeled dataset.

Alternatively, one can consider different cues successively instead of in parallel [25,38]. In [38], e.g., first a coarse segmentation of the moving objects is obtained from using motion information, and then the segmentation is refined by considering the local image gradient. In [25], the authors deal with colour and texture images by first quantizing the image in the colour space, and then introducing a spatial criterion to obtain the segmentation. We refer to Clark and Yuille [21] for a classification of the different approaches in cues integration.

The present article comprises and extends work that has been presented at conferences. The variational framework based on level sets has been presented in [55]. In [54] it has been applied to texture segmentation based on the nonlinear structure tensor. Finally, in [10], colour, texture, and motion information have been integrated. In particular, the present paper extends these works by: (i) a detailed investigation of feature coupling, including the way of normalisation, (ii) by showing that with a Gaussian region model, the partitioning becomes independent from the contrast of the features and comprises an implicit weighting of features by their discriminative power, (iii) by embedding our method in a multi-scale setting, and finally (iv) by a significantly extended experimental evaluation.

1.2. Paper organisation

In the following section, the extraction of the texture and motion features will be presented. This also comprises the combination of all features in a single feature vector. After that, in Section 3, the variational formulation for image partitioning based on level set methods will be introduced. Section 4 shows experiments on image segmentation using different feature combinations. Finally, Section 5 introduces an extension that allows the tracking of multiple objects. Experimental results are shown for this application as well. The paper is concluded by a summary.

2. Feature extraction and enhancement

2.1. Texture features from a nonlinear structure tensor

The nonlinear structure tensor introduced in [14,70] is based on the classic linear structure tensor [6,30,53]

$$J_\rho = K_\rho * (\nabla I \nabla I^\top) = \begin{pmatrix} K_\rho * I_x^2 & K_\rho * (I_x I_y) \\ K_\rho * (I_x I_y) & K_\rho * I_y^2 \end{pmatrix}, \quad (1)$$

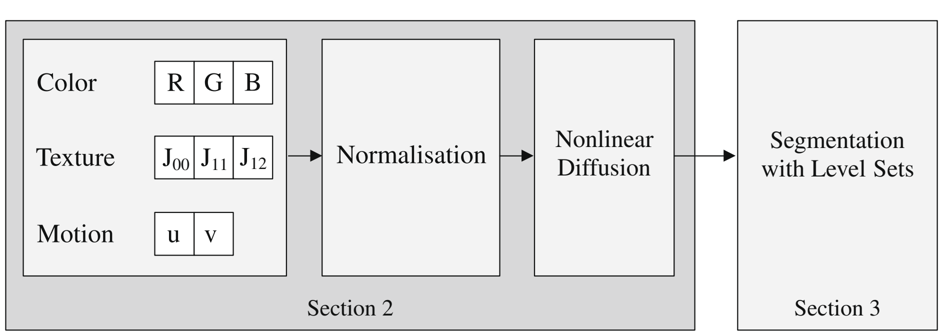

Fig. 1. Summary of our approach.