Multi-class Video Objects Segmentation Based on Conditional Random Fields

1, 2 Zhiwei HE, 1 Lijun XU, 1 Wei ZHAO, 1 Mingyu GAO
1 College of Electronic and Information Engineering, Hangzhou Dianzzi University, Xiasha Higher Education Zone, 310018, China
2 School of Information Engineering, Hangzhou Dianzzi University, Xiasha Higher Education Zone, 310018, China
1 Tel.: (86) 0571-86919153, fax: (86) 0571-86919153
1 E-mail: zwhe@hdu.edu.cn

Received: 22 October 2013 / Accepted: 9 January 2014 / Published: 31 January 2014

Abstract: Video object segmentation has been widely used in many fields. A conditional random fields (CRF) model is proposed to achieve accurate multi-class segmentation of video objects in the complex environment. By using CRF, the color, texture, motion characteristics and neighborhood relations of objects are modeled to construct the corresponding energy functions in both the temporal and spatial domains. The model is trained with annotated samples by using LogitBoost classifier. The energy function is amended by adding a constraint factor which is used to indicate the interaction between two adjacent images in the video sequence. Experimental results show that the proposed algorithm can achieve high performance for multi-class objects segmentation in videos under complex environment. It can also get good recognition results when dealing with multi-viewed objects or serious sheltered objects. Copyright © 2014 IFSA Publishing, S. L.

Keywords: Video object segmentation, Conditional random fields model, LogitBoost classifier, Constraint factor.

1. Introduction

In recent years, video object segmentation algorithm is widely applied in intelligent transportation systems [1], video surveillance [2], human-computer interaction systems [3] and other fields. It has become a hot research area in the field of computer vision. The effectiveness of video object segmentation algorithm is the key to the success of subsequent application. An accurate and efficient segmentation algorithm can greatly reduce the difficulty of subsequent process. Therefore, designing an accurate and efficient segmentation algorithm is a key issue.

For single-class object segmentation algorithm, scholars in the world wide have conducted a lot of researches and achieved good results. To model the video sequence’s background, Yang et al. [4] proposed a method which employs the method of background subtraction, which is simple and easy to be fulfilled, but the algorithm doesn’t work well for dynamic background; while, Stauffer et al [5] put forward the Gaussian mixture model, which can deal with different backgrounds, either indoor or outdoor. Wang et al. [6] applied the CRF model to segment video objects, which models the neighborhood information of the video sequence in the spatial-temporal fields simultaneously, considering the
correlation between two adjacent images, and a filter is constructed which can be recursively updated according to their correlation. Yiping Chu et al [7] improved the CRF model according to a series of empirical values, and proposed an adaptive CRF-based segmentation algorithm. The algorithm can calculate the characteristic intensity of conditional random fields adaptively, and it can approximate the result of segmentation with optimal parameters.

Compared to the object segmentation algorithm for single class, segmentation with multi-class object should model different characteristics among all the classes. At the same time, it should model constrained interrelations between classes (such as fish live in the water, the sky is above the grass, etc). Thus, each region of these segmented images has different semantic category labels. A CRF based multi-class object segmentation algorithm is proposed in this paper. The algorithm can achieve high performance for multi-class object segmentation in videos under complex environment. It can also get good recognition results when dealing with multi-viewed and serious sheltered objects.

2. CRF-Based Multi-Class Object Segmentation Model

2.1. CRF Model

Conditional Random Fields (CRF) model is a probability model based on undirected graphs. By modeling the images’ color, texture, motion characteristics and field relations, it can construct object segmentation model successfully. With the optimized iterative algorithm, it is then able to segment objects successfully.

Suppose that random variable \( X \) is a sampling value of the video frames' pixels in the RGB color space, and random variable \( L \) represents a classification label of \( X \). Given the random variable \( X \), its conditional probability of classification label set is defined as a two-dimensional CRF model.

\[
P(L \mid X) = \frac{1}{Z(X)} \exp(-E(L;X)), \quad (1)
\]

\[
E(L;X) = U(L;X) + V(L;X), \quad (2)
\]

\[
Z(X) = \sum_L \exp(-E(L;X)), \quad (3)
\]

where \( Z(X) \) is the potential function, and can get a normalized result for the required conditional probability; \( E(L;X) \) is the energy function; \( U(L;X) \) represents the local energy function, including the color, texture, location and other features of a single pixel; \( V(L;X) \) is the neighborhood energy function, including the time and spatial neighborhood.

Based on the CRF model, we can get segmentation results of images according to the following form.

\[
\hat{L} = \arg \max_L P(L;X), \quad (4)
\]

2.2. Construction of Energy Function

For the CRF-based object segmentation model above, the segmentation results mainly depend on the definition of the energy functions. The energy functions here include local energy function and neighborhood energy function.

Local energy function \( U(L;X) \) can be obtained by the sum of each pixel's characteristic function. Considering that the surrounding pixels have certain influence on the current pixel, we define a neighborhood energy function \( V(L;X) \) which mainly represents the influence of its 8 neighbor pixels on the current pixel. Therefore, the energy function can be defined as the following formula.

\[
E(L;X) = \sum_{i=1}^{a} v_{i}(L_{i};X) + \sum_{i,j \in \partial} v_{ij}(L_i;L_j;X), \quad (5)
\]

2.2.1. Construction of Local Energy Function

Local energy function models a single pixel's color, position, texture and other characteristics.

In this paper, local energy is depicted with different image features computed as follows. Firstly, the three color information R, G, B in each color channel are used as three different features. Secondly, the 17-dimensional filter bank proposed by J. Winn et al [12] is applied to compute the images' texture features: three Gaussian filters with kernels \( \sigma = 1, 2, 4 \) are applied to each CIE L, a, b channel [13] to get 9 responses; four LoG filters [14] with kernels \( \sigma = 1, 2, 4, 8 \) are applied to the L channel to produce 4 responses; and four derivatives of Gaussians, divided into two \( x \)- and \( y \)-aligned sets and each with two different \( \sigma = 2, 4 \), are applied to the \( L \) channel to get 4 responses; therefore, each pixel in each image has associated with a 17-dimensional feature vector. Thirdly, the nine HOG [15] features at nine different orientations are also appended to the feature vector. Next, a \( 3 \times 3 \) grid of cells around each pixel is defined, and the mean and standard deviation of the aforementioned features (29-dimensional) in each grid is computed. Finally, the raw image features, the mean and standard deviation features and the
normalized $x$ and $y$ location of the pixel are appended together into a local feature vector. As the input of LogitBoost classifier, the feature matrix and different label categories can be used to estimate the probability of each label with different features. The concrete steps of LogitBoost algorithm was shown in [8] as follows

$$P(L_e = l | X_e) = \frac{\exp(F_{il})}{\sum_{i=1}^{m} \exp(F_{il})}, \quad (6)$$

where $F_{il}$ denotes the output of the final classifier, and the local energy of each pixel is the logarithm of the conditional probability obtained above,

$$\psi(L_e; X_e) = -\log P(l | X_e), \quad (7)$$

2.2.2. Construction of Spatial Domain Energy

Spatial neighborhood energy function reflects the spatial intensity relationship among pixels in the image. Considering the interaction of each pixel within an 8-neighbor area, we define the spatial domain energy of two adjacent pixels as follows:

If $L_i \neq L_j$, then

$$\psi_{ij}(L_i, L_j; X) = \frac{\lambda_i}{d_{ij}} \exp \left( - \frac{\|x_i - x_j\|}{2d_{ij}} \right), \quad (8)$$

where $x_i$ and $x_j$ represent pixel values of two corresponding adjacent pixels respectively; $\|x_i - x_j\|$ is the expectation of intensity difference between adjacent pixels; $d_{ij}$ denotes the distance between two adjacent pixels, for pixels in 4-neighbor, $d_{ij} = 1$ ; while for the diagonal pixels, $d_{ij} = \sqrt{2}$ ; the non-negative parameters $\lambda_i$ is obtained by cross-validation with the training images.

According to (8), the value of $\psi_{ij}$ mainly depends on the difference of $x_i$ and $x_j$. If the difference is smaller, the neighborhood pixels will make more contributions to the neighborhood energy; if larger, contributions will be less.

2.3. Inference of the Model

According to the formula $\hat{L} = \arg \max_{L} P(L; X)$, we can get the CRF-based segmentation results of images. This is actually a MAP problem. Usually it can be solved by the Gibbs sampling algorithm, the iterative conditional model, or the belief propagation algorithm.

Because the Gibbs sampling algorithm can take advantage of the coherence information between video frames effectively, and can greatly reduce the iteration time and computation cost, we use the Gibbs sampling algorithm to solve the MAP problem. The initial segmentation of the current frame is taken as an initial input to the Gibbs sampling algorithm, and the convergence of the total energy can be achieved through continuous iterations.

2.4. Experimental Results on Static Images

The experiments for static images are done on images in the MSRC-21 class database [11].

The MSRC-21 class database is one of the most complex and comprehensive annotated databases. It consists of 591 images and is labeled with 21 classes of objects. These images contain different viewed, as well as sheltered objects with high structural characteristics and high texture features.

Fig. 1 shows some sample images along with the corresponding annotation in the MSRC-21 class database, with different color markings represent different classes.

With the method proposed in this paper, we first train the CRF model with the annotated images in the MSRC-21 database and obtain the parameters of the model. The trained model with learnt parameters is then applied to segment the objects in new images. Some of the segmentation results are shown in Fig. 2.

Pictures in Fig. 2(a) are the original images in the MSRC-21 class database. Fig. 2(b) are the segmentation results by taking only local energy function into consideration, while the energy function in the spatial neighborhood are not modeled. Fig. 2(c) shows the segmentation results obtained by employing both the local energy and the spatial neighbour energy.

As can be seen from the figure, the majority of pixels in the images can be successfully recognized and segmented, but obviously, the results in Fig. 2(c) are better than Fig. 2(b).

In Fig. 2(b), some pixels are misclassified as the spatial information between neighbor pixels are discarded.

When the spatial energy is applied, good recognition results can be achieved as seen from Fig. 2(c).

On the other hand, as seen from Fig. 2(a), the feet and some parts of the body of the sheep are sheltered by the grass seriously, but it can still be segmented properly according to the proposed algorithm. Obviously, this algorithm can be applied to seriously sheltered objects.
**Fig. 1.** Example images with corresponding annotation graphs from MSRC-21 database.

<table>
<thead>
<tr>
<th>Object classes</th>
<th>Building</th>
<th>Grass</th>
<th>Tree</th>
<th>Cow</th>
<th>Sheep</th>
<th>Sky</th>
<th>Aeroplane</th>
<th>Water</th>
<th>Face</th>
<th>Car</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bike</td>
<td>Flower</td>
<td>Sign</td>
<td>Bird</td>
<td>Book</td>
<td>Chair</td>
<td>Road</td>
<td>Cat</td>
<td>Dog</td>
<td>Body</td>
</tr>
</tbody>
</table>

(a.1)                                   (b.1)                                                                 (c.1)

(a.2)                                   (b.2)                                                                 (c.2)

(a.3)                                   (b.3)                                                                 (c.3)

**Fig. 2.** (a) The original picture (b) CRF-based multi-class object segmentation results considering only the local energy function (c) Segmentation results by the proposed method.
3. The CRF-Based Multi-Class Segmentation Model in Video Sequences

3.1. The Improved Energy Function

In a video sequence, considering that there must be some correlation between corresponding pixels in two adjacent images, the energy function is improved with a constraint factor added \([9]\). The constraint factor represents the energy relationship between two pixels in the current frame and the previous frame. It is expressed in the following formula.

\[
E(L_1; L_2; X_1; X_2) = E(L_2; X_2) + E(L_1; X_1) + \sum_{(p,q) \in P} \psi_{pq}(L_{1,p}, L_{2,p}; X_1; X_2) ,
\]

where \(X_1\) and \(X_2\) are two adjacent frames in video sequences; and \(P = \{ (p_1, p_2) : p_1 \in X_1, p_2 \in X_2 \}\) is a set of pairs of corresponding pixels. \(E(L_1; X_1)\), \(E(L_2; X_2)\) represent the sum of the local energy function and the spatial neighborhood energy function of video frames \(X_1\), \(X_2\) respectively; while, the last term in (9) expresses the constraint between pixel \(p\) in the current video frame and the corresponding pixel \(q\) in the previous video frame. The constraint function is defined as follows:

\[
\psi_{pq}(L_{1,p}, L_{2,p}; X_1; X_2) = \begin{cases} \lambda_C C_{pq} & L_{1,p} \neq L_{2,p} \\ 0 & \text{otherwise} \end{cases} ,
\]

where \(L_{1,p}\) and \(L_{2,p}\) respectively denote the label value of pixel \(p\) in the current video frame and pixel \(q\) in the previous video frame. If they match, the constraint factor is \(0\); otherwise it is not \(0\). As there will be some differences during the matching process, \(C_{pq}\) is introduced to indicate the accuracy of the matching when necessary, whose definition can be find later in (12); \(\lambda_C\) is the constant, which can also be obtained by cross validation.

3.1.1. Hierarchical Cluster Analysis

Let \(g_i\) denote a scene descriptor of image \(i\) in the image set \(X\) (training or test set); the equation, expressed as \(D_0 = \|g_i - g_j\|^2\) is the distance between the scene descriptor of image \(i\) and the scene descriptor of image \(j\); \(N_{\text{max}}\) indicates the allowable maximum number of images in each class.

Firstly, we make cluster analysis for each image. some of the two classes are then merged for several times until the number of images is larger than \(N_{\text{max}}\).

In each iteration process, we must find two clusters which obey the follow two rules: (1) the number of images in each class must be smaller than \(N_{\text{max}}\); (2) the distance between them is less than the minimum distance between their elements. In other words, clusters \(X_a\) and \(X_b\) need meet the following requirement.

\[
(a,b) = \arg \min_{a \neq b \in \{1, \ldots, \ell \}} \{ \min_{a \neq b \in \{1, \ldots, \ell \}} D_0 \} ,
\]

Then the two clusters satisfying the requirement are merged together, and a new cluster \(X_a \cup X_b\) is obtained.

3.1.2. Image Matching

In order to find similar regions in the same cluster, the patch-match algorithm \([10]\) is employed in this paper. It is an approximation algorithm. It firstly segments one image into many small pieces, and then compares small pieces in one image with those small pieces in another image. Finally, the most similar small pieces within a certain distance are matched. The detailed process can be divided into three steps:
- Initialize the nearest neighbor field (NNF) with random offsets;
- Suppose that the image is naturally smooth, then take a good match to the neighboring small piece;
- Set the best offset as the radius, then explore small pieces randomly and find a better match.

Repeat the second and third steps, until it converges or reaches the maximum iteration number.

In the process of matching, some image blocks may not find the best matching. In order to make the image blocks, which matched no so good, be applicable to this model, a parameter \(C_{pq}\) is introduced to represent the quality of the matching. Using the patch-match algorithm, the results obtained by matching pixel \(p\) of video frame \(X_1\) to pixel \(q\) of video frame \(X_2\) are denoted by \(S_{pq}\), and then \(C_{pq}\) can be expressed as

\[
C_{pq} = \exp\{-S_{pq}/2\beta\} ,
\]

where \(\beta\) designates the mean of matching results obtained by employing patch-match algorithm to all of the adjacent video frames in the video sequence.

3.2. Results and Discussion

In order to test the segmentation results of the CRF-based multi-class segmentation model in video sequences, multi-class segmentation experiments have been made on videos downloaded online. The experimental results are shown in Fig. 3.
Fig. 3 (a). The 33rd image.

Fig. 3 (g). The segment result of the 33rd image.

Fig. 3 (b). The 156th image.

Fig. 3 (h). The segment result of the 156th image.

Fig. 3 (c). The 205th image.

Fig. 3 (i). The segment result of the 205th image.

Fig. 3 (d). The 431st image.

Fig. 3 (j). The segment result of the 431st image.
As shown in Fig. 3, in which Fig. 3(a) – Fig. 3(f) respectively represent the 33rd, 156th, 205th, 431st, 471st, and 491st frame in the video sequence, and Fig. 3(g) – Fig. 3(l) are the corresponding segmentation results. Although some pixels are misclassified in the segmentation process, from the segmentation results, we can see that most of the pixels are classified correctly, and the grassland, cattle, water have been segmented accurately, indicating that the algorithm has good classification effects.

From Fig. 3(c), we can see cattle can been accurately identified, although they are far away and can't be seen clearly. It shows that the algorithm can also segment objects seriously sheltered or from different angles in video sequences. Meanwhile, it can significantly shorten the segmentation time and improve efficiency by matching.

4. Conclusions

Most traditional CRF-based multi-class video object segmentation algorithm models take only the local energy function and the spatial neighborhood energy function into consideration for building the model, and the objects in images are segmented separately, while discarding the regional similarity between two adjacent images in video sequences. The algorithm in this paper adds a constraint factor when modeling the energy function. The factor represents interactions between two adjacent images, and employs contextual constraints in video sequences. Experimental results show that, under a complex environment, the algorithm can achieve good segmentation results for video objects, even if the objects are severe sheltered or in different views.

Acknowledgements

This work was supported by the Nature Science Foundation of China (NSFC) Grant #61001216, Zhejiang Provincical Natural Science Foundation of China Grant #Z1110741, #LQ13F010011, Zhejiang Educational Committee Foundation of China Grant #Z201122745.

References


