ESTIMATING ADAPTIVE COEFFICIENTS OF EVOLVING GMMS FOR ONLINE VIDEO SEGMENTATION

Ioannis Kaloskampis and Yulia A. Hicks

School of Engineering, Cardiff University, UK

ABSTRACT

A new, online, evolving video segmentation algorithm is presented in this paper. The proposed method segments each video frame using an evolving Gaussian mixture model (GMM) whose adaptive coefficient is automatically adjusted to cater for abrupt changes between consecutive frames.

The proposed method is tested against another algorithm, which keeps the adaptive coefficient constant. The comparison shows the advantage of altering the value of the adaptive coefficient according to change in the scene.

Index Terms— Computer vision, video segmentation, model adaptation, Gaussian mixture, on-line processing

1. INTRODUCTION

Spatial video segmentation aims to divide each frame into meaningful area segments. With video being one of the fastest growing resources of data, resulting from the popularisation of video surveillance and the vast increase of video content on the web, multimedia applications featuring retrieval of meaningful objects and detection of scenes are necessary for efficiently processing such large amounts of information. Thus, spatial video segmentation is a prominent research area.

Compared to single image segmentation, video segmentation brings in the additional challenge of catering for segmentation consistency throughout the image sequence. Towards this effort, researchers follow two main approaches [1]: the first is tracking of regions from frame to frame and the second is applying spatio-temporal grouping techniques [2, 3, 4]. The algorithm proposed in this article falls in the first category.

When tracking regions from frame to frame, segmentation consistency is achieved by using prior knowledge stored in a model [5]. Two representative algorithms following this principle are Kohli and Torr [6], where user specified segmentation cues on the first frame of the video sequence are used to ensure consistency, and Goldberger and Greenspan [1] where the feature pdf obtained from the previous frame is modelled as a GMM and used to estimate the pdf of the current frame using maximum *a posteriori* learning. However in both algorithms segmentation results rely on the correct selection of the number of clustering components. This limitation is not critical for videos featuring a fixed number of important objects, captured with a static camera (*e.g.* news broadcast); on the other hand, for videos captured with moving cameras, with objects entering and exiting the scene, the number of clustering components is variable. Charron and Hicks [5] address this problem by proposing an algorithm which automatically changes the parameters of the GMM including the number of components over time according to changes from frame to frame in the video, which leads to an evolving model. The ratio of the contribution of old and new information in the evolving model is kept constant throughout the video sequence.

In this paper a new video segmentation algorithm is presented, which builds on the idea of varying the parameters of the model used for segmentation according to difference from frame to frame. The novelty of the proposed algorithm is in automatically adjusting the contribution of old and new information in the evolving model according to the measured difference between consecutive frames. We show in the results section that the proposed method offers more uniform segmentation results in scenes featuring significant changes.

The rest of the paper is structured as follows. The proposed algorithm is first described in section 2. Experimental results for a publicly available video are presented in section 3. Conclusions are presented in section 4.

2. METHODOLOGY

In this section the novel evolving video segmentation algorithm is described. The sequence of frames to be segmented is represented as $\{I^{(t)}\}_{t=1,\ldots,T}$. In a standard approach to segment an image using a GMM, first a feature vector is extracted for each pixel in the frame. The pdf of these feature vectors, $p(\mathbf{x}|\theta)$, is modelled using a GMM:

$$p(\mathbf{x}|\theta) = \sum_{i=1}^{K} a_i \mathcal{G}(\mathbf{x}|\mathbf{m}_i, \mathbf{C}_i), \tag{1}$$

where $\mathbf{x}$ the feature vector, K the number of GMM components which varies from frame to frame, $a_i, i = 1, \ldots, K$

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/K014307/1 and the MOD University Defence Research Collaboration in Signal Processing.

are the mixture weights and $\mathcal{G}(\mathbf{x}|\mathbf{m}_i, \mathbf{C}_i)$ are the component Gaussian densities with mean vector $\mathbf{m}_i$ and covariance matrix $\mathbf{C}_i$. It is assumed that each Gaussian in the mixture corresponds to a segment of the frame. To segment an image each pixel in the frame is attributed to a segment according to its probability as estimated with the pdf.

When expanding to online video segmentation, the parameters of the GMMs representing segments including the number of Gaussians should change over time according to changes in the video leading to an evolving GMM. An approach followed in this article achieves the update of GMM parameters by merging the evolving GMM with a temporary GMM trained on the current frame, $I^{(t)}$ [5]. After merging, the temporary GMM and all data related to it (image, extracted features) are discarded as the algorithm only requires the parameters of the evolving GMM for the next step; this results in efficient handling of computer memory and storage.

The contribution of the evolving GMM, $\theta^{(t)}$, and the temporary GMM, θ' in the merged model, $\theta^{(t+1)}$ is determined by an adaptive coefficient, c. In [5] coeffcient c was set to a specific value, ensuring equal contribution from the two GMMs, and kept constant throughout the image sequence. This proves sufficient when the difference from frame to frame is not significant. To improve segmentation accuracy in videos featuring important changes from frame to frame, an algorithm is proposed in this paper which detects them and adjusts c according to the significance of these changes. Thus, in our work c is variable and estimated by computing change from frame to frame.

In summary, components are merged as follows. First, the mixture components from the two GMMs are concatenated (section 2.1). To evaluate the adaptive coefficient, difference between the current and previous frame is estimated using histogram differences (section 2.2). Finally, the components are merged with the expectation-maximisation (EM) algorithm (section 2.3). Detailed description of the algorithm follows.

2.1. GMM concatenation

The components of GMMs θ' and $\theta^{(t)}$ are concatenated yielding an intermediate GMM, denoted as θ'':

$$\begin{aligned}\theta'' = \{&ca'_1, \ldots, ca'_{K'}, (1-c)a_1^{(t)}, \ldots, (1-c)a_{K^{(t)}}^{(t)},\\ &\mathbf{m}'_1, \ldots, \mathbf{m}'_{K'}, \mathbf{m}_1^{(t)}, \ldots, \mathbf{m}_{K^{(t)}}^{(t)},\\ &\mathbf{C}'_1, \ldots, \mathbf{C}'_{K'}, \mathbf{C}_1^{(t)}, \ldots, \mathbf{C}_{K^{(t)}}^{(t)}\}\end{aligned} \quad , \quad (2)$$

where $c \in [0, 1]$; as c tends towards 1, contribution of the evolving GMM increases and contribution of temporary GMM decreases. Thus, information collected from past frames is taken more into account, which is desirable for small changes from frame to frame. The impact of the temporary GMM increases as c tends towards 0; in this case new information, collected from current frame is favoured, which is desirable for abrupt changes. The next subsection proposes an algorithm which adjusts c according to change detected from frame to frame.

2.2. Estimating the adaptive coefficient

The difference between two consecutive frames is quantified by representing each frame with a suitable descriptor and measuring distances between descriptors using a metric. Regarding image representation, as colour varies over an image or image part, appearance is best described by the distribution of features rather than by individual feature vector [7]. Thus, each frame is represented with its colour histogram. Distance between histograms is computed with the Minkowski distance. This metric combines relatively low complexity, which is important for on-line video segmentation and good discriminative accuracy [7]. For two images, X and Y, represented by colour histograms $f(i; X)$ and $f(i; Y)$, with equal number of bins l and $i \in \{1, \ldots, l\}$ the bin index, the Minkowski distance, $\mathcal{L}_p$ is defined as:

$$D(X, Y) = \left(\sum_i |f(i; X) - f(i; Y)|^p \right)^{1/p} , \quad (3)$$

with p the order of the distance. Based on results of [7], $p = 1$ is chosen *i.e.* the first order Minkowski distance, $\mathcal{L}_1$.

The adaptive coefficient c is defined to be correlated to the estimated difference between two consecutive frames, $\mathcal{L}_1(t)$. A linear relationship between c and $\mathcal{L}_1(t)$ is assumed. Thus, c is given by the equation:

$$c = \begin{cases} 0, & v < 0 \\ \beta_0 + \beta_1 \mathcal{L}_1(t), & 0 \leq v \leq 1 \\ 1, & v > 1 \end{cases} \quad (4)$$

where $v = \beta_0 + \beta_1 \mathcal{L}_1(t)$. The values of the coefficients β_0 and β_1 are estimated on the basis of a training set, N_{train}, comprising a number of video frames. It is assumed that there are no abrupt changes in this set. For the selected frames, average, minimum and maximum value of Minkowski distance are estimated and denoted as $\mathcal{L}_{1,avg}$, $\mathcal{L}_{1,min}$ and $\mathcal{L}_{1,max}$ respectively. A dataset $\{c_i, \mathcal{L}_{1,i}\}_{n=1}^3$ is formed with:

$$\begin{aligned} \{c_1, \mathcal{L}_{1,1}\} &= \{\bar{c}, \mathcal{L}_{1,avg}\} \\ \{c_2, \mathcal{L}_{1,2}\} &= \{\bar{c} + s, \mathcal{L}_{1,min}\} \\ \{c_3, \mathcal{L}_{1,3}\} &= \{\bar{c} - s, \mathcal{L}_{1,max}\} \end{aligned} \quad (5)$$

In (5), $\bar{c}$ is the starting value of the adaptive coefficient, c and s its deviation in the training set. Both are empirically initialised according to the content of the training sequence N_{train}. They satisfy the constraints $\bar{c} \in [0, 1]$ and $\bar{c} \pm s \in [0, 1]$ so that $c \in [0, 1]$ as stated in section 2.1. In the same section it is specified that as change between consecutive frames

increases, c decreases. Change in this work is measured using $\mathcal{L}_1$; thus, as it is $\mathcal{L}_{1,min} < \mathcal{L}_{1,avg} < \mathcal{L}_{1,max}$ it is also $\bar{c} - s < \bar{c} < \bar{c} + s$. Therefore $s > 0$ holds. The regression coefficients β_0 and β_1 are estimated using simple linear regression on the dataset defined in (5) as illustrated, for example, in [8].

2.3. Merging GMMs

Having estimated c, the concatenated model obtained by (2) is merged in a concise form using the EM algorithm, as in Charron and Hicks [5]. A GMM, θ'' of $K' + K^{(t)}$ components is converted to a new, concise GMM, $\theta^{(t+1)}$ of $K^{(t+1)}$ components, with $K^{(t+1)} \leq K' + K^{(t)}$. Estimation of parameters of $\theta^{(t+1)}$, *i.e.* its priors $\{a_j^{(t+1)}\}_{j=1,\ldots,K^{(t+1)}}$, means $\{\mathbf{m}_j^{(t+1)}\}_{j=1,\ldots,K^{(t+1)}}$ and covariances $\{\mathbf{C}_j^{(t+1)}\}_{j=1,\ldots,K^{(t+1)}}$ is now described, setting $t+1=\tau$. The E-step estimates the component responsibilities, w_{ij}:

$$w_{ij} = \frac{\left[\mathcal{G}(\mathbf{m}_i''|\mathbf{m}_j^{(\tau)}, \mathbf{C}_j^{(\tau)}) e^{-\frac{1}{2}tr((\mathbf{C}_j^{(\tau)})^{-1}\mathbf{C}_i'')}\right]^{M_i} a_j^{(\tau)}}{\sum_{k=1}^{K^{(\tau)}} \left[\mathcal{G}(\mathbf{m}_i''|\mathbf{m}_k^{(\tau)}, \mathbf{C}_k^{(\tau)}) e^{-\frac{1}{2}tr((\mathbf{C}_k^{(\tau)})^{-1}\mathbf{C}_i'')}\right]^{M_i} a_k^{(\tau)}} \tag{6}$$

In (6), $\mathcal{G}(\mathbf{x}|\mathbf{m}, \mathbf{C})$ is a Gaussian with mean $\mathbf{m}$ and covariance $\mathbf{C}$. With M_i the number of samples attributed to the i^{th} component of the old model is denoted.

During the M-step, the likelihood of the new model, $\theta^{(\tau)}$ given the old model $\theta^{(t)}$ is maximised as follows [5]:

$$\hat{a}_j^{(\tau)} = c\sum_{i=1}^{K'} w_{ij} a_i' + (1-c)\sum_{i=K'+1}^{K'+K^{(t)}} w_{ij} a_{i-K'}^{(t)} \tag{7}$$

$$\hat{\mathbf{m}}_j^{(\tau)} = c\sum_{i=1}^{K'} w_{ij} \mathbf{m}_i' + (1-c)\sum_{i=K'+1}^{K'+K^{(t)}} w_{ij} \mathbf{m}_{i-K'}^{(t)} \tag{8}$$

$$\begin{aligned}\hat{\mathbf{C}}_j'^{(\tau)} =& c\sum_{i=1}^{K'} w_{ij}(\mathbf{C}_i' + \mathbf{m}_i'\mathbf{m}_i'^T) \\ &+ (1-c)\sum_{i=K'+1}^{K'+K^{(t)}} w_{ij}(\mathbf{C}_{i-K'}^{(t)} + \mathbf{m}_{i-K'}^{(t)}\mathbf{m}_{i-K'}^{(t)T}) \\ &- \hat{\mathbf{m}}_j^{(\tau)}\hat{\mathbf{m}}_j^{(\tau)T}\end{aligned} \tag{9}$$

In (7, 8, 9), the current estimate is denoted as $\hat{\theta}^{(\tau)}$. The EM steps are repeated until an arbitrary threshold is reached.

3. EXPERIMENTAL RESULTS

In this section experimental results of the proposed method are presented and compared against those obtained by the method described in [5]. Both methods are used to segment the publicly available Flower Garden video sequence [9]. Application of video segmentation on this video is challenging, as it is captured using a moving camera, which results in new objects entering the scene.

Each video frame is converted into a set of features. This is achieved by representing each pixel with a three-dimensional colour descriptor in the $L*a*b*$ colour space, which was shown to be approximately perceptually uniform [1]. The position of the pixel, as defined by its spatial coordinates (x, y), is also included in its vector.

Colour histograms are constructed for each frame using 10 bins for each colour channel, resulting in histograms of 10^3 bins. The first 50 frames of the video are used for training. Following section 2.2, average, minimum and maximum values of $\mathcal{L}_1$ between histograms of consecutive frames are computed for these 50 frames. Then, the dataset (5) is formed using these values. Parameters $\bar{c}$ and s for the dataset are initialised to $\bar{c} = 0.7$ and $s = 0.03$. Application of linear regression on the formed dataset yields an equation correlating the adaptive coefficient c and $\mathcal{L}_1(t)$. Thus c can be computed for the whole sequence (Fig. 1). Abrupt decrease of c is observed in the frame interval 182-192 which coincides with the event of a large object (tree) entering the scene. As discussed in section 2.1, for abrupt changes c should decrease, so that new information is favoured. It is observed that the proposed algorithm displays the desirable behaviour of detecting change in the scene and decreasing the value of c accordingly.

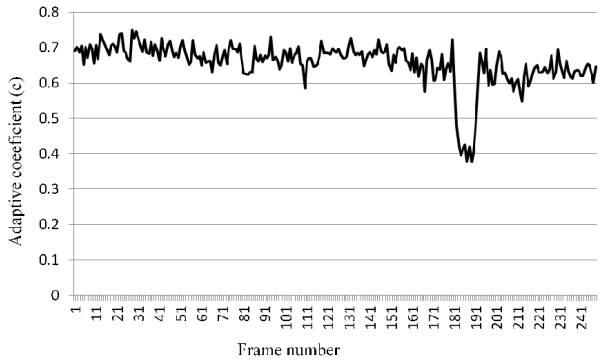
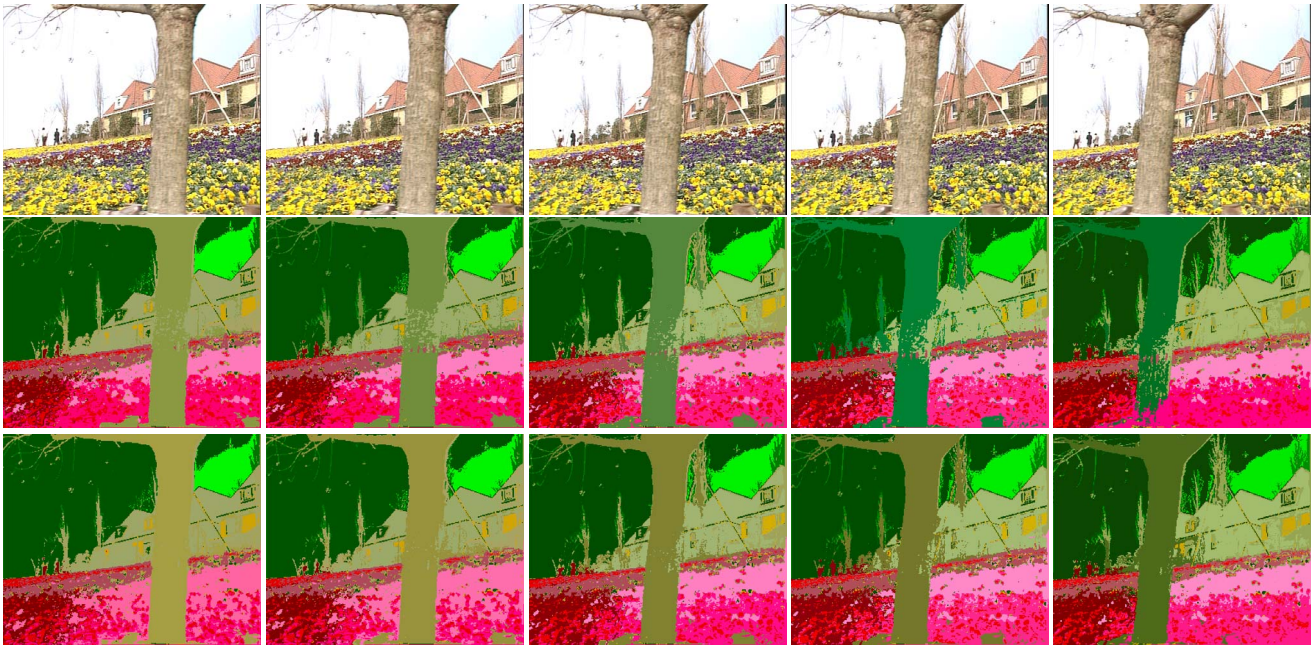


Fig. 1. Fluctuation of adaptive coefficient (c) within the flower garden video sequence.

Qualitative segmentation results of the proposed method are presented in Fig. 2 and compared against those obtained by the algorithm of [5] for five frames. The algorithm proposed in this work segments the object entering the scene after frame 182 (tree) more uniformly than the algorithm of [5].



(a) Frame 208 (b) Frame 211 (c) Frame 216 (d) Frame 219 (e) Frame 225

Fig. 2. Segmentation results. Top row: original image; middle row: algorithm [5]; bottom row: our method.

4. CONCLUSION

A new, online, evolving video segmentation algorithm was presented in this paper. The proposed method segments each video frame using an evolving GMM whose parameters, inclusive of the number of components, are automatically adjusted to cater for abrupt changes between consecutive frames.

The proposed algorithm was tested in a publicly available video sequence featuring camera motion, resulting in significant change from frame to frame. It was demonstrated that altering the value of the adaptive coefficient according to change measured from frame to frame results in better segmentation of the scene. This was shown by comparing the proposed method with a previous one, where the adaptive coefficient is constant throughout the video sequence.

5. REFERENCES

[1] J. Goldberger and H. Greenspan, "Context-based segmentation of image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 463–468, 2006.

[2] R. Megret and D. Dementhon, "A survey of spatio-temporal grouping techniques," Tech. Rep., UMIACS-2002-83, 2002.

[3] Yuchi Huang, Qingshan Liu, and D. Metaxas, "Video object segmentation by hypergraph cut," in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2009, pp. 1738–1745.

[4] J. Lezama, K. Alahari, J. Sivic, and I. Laptev, "Track to the future: Spatio-temporal video segmentation with long-range motion cues," in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[5] C. Charron and Y. Hicks, "An evolving MoG for online image sequence segmentation," in *17th IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 2189–2192.

[6] P. Kohli and P. H S Torr, "Dynamic graph cuts for efficient inference in markov random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2079–2088, 2007.

[7] Y. Rubner, J. Puzicha, C. Tomasi, and J. M. Buhmann, "Empirical evaluation of dissimilarity measures for color and texture," *Computer Vision and Image Understanding*, vol. 84, no. 1, pp. 25–43, Oct. 2001.

[8] J. Fox, *Applied regression analysis, linear models, and related methods*, Sage Publications, Thousand Oaks, Calif., 1997.

[9] Xiph.Org Foundation, "Xiph.org test media," `http://media.xiph.org/`, 2013, [Online; access. 10 Dec. 2013].