On fusion for robust motion segmentation

Conference or Workshop Item

Accepted Version


It is advisable to refer to the publisher’s version if you intend to cite from the work. See Guidance on citing.
Published version at: http://dx.doi.org/10.1109/AVSS.2015.7301776

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur
On Fusion for Robust Motion Segmentation

Longzhen Li, Anna Ellis and James Ferryman
Computational Vision Group, School of Systems Engineering, University of Reading, UK
{ longzhen.li | a.l.ellis | j.m.ferryman }@reading.ac.uk

Abstract

While a multitude of motion segmentation algorithms have been presented in the literature, there has not been an objective assessment of different approaches to fusing their outputs. This paper investigates the application of 4 different fusion schemes to the outputs of 3 probabilistic pixel-level segmentation algorithms. We performed an extensive experimentation using 6 challenge categories from the changedetection.net dataset demonstrating that in general simple majority vote proves to be more effective than more complex fusion schemes.

1. Introduction

Motion segmentation represents one of the most important low-level tasks in visual surveillance. However, to date, no single segmentation algorithm is robust under the wide range of environmental and other challenges present in surveillance footage. It is conjectured that fusion of the outputs of multiple algorithms’ outputs operating on the same data will enhance robustness. While some recent published segmentation methods implicitly include fusion schemes within their overall approaches, including exploitation of different feature types [9, 10, 14], the merits of the different fusion strategies themselves has not been explicitly studied in the literature. Furthermore in initiatives such as the CDnet challenge [15], which have addressed the need to evaluate change and motion detection approaches providing a benchmark dataset and website hosting results, only one fusion approach (majority vote, applied to the top three independently performing algorithms as well as applied to all ranked algorithms) is explicitly considered and included in the published ranked list. This work addresses this issue by quantitatively evaluating 4 different fusion approaches applied to 3 pixel-based probabilistic algorithms operating on 6 categories from the CDnet challenge.

2. Related Work

Prior work for fusing pixel-level segmentation algorithms include application of Expectation Maximisation (EM), Bayes and Boosting amongst others. [16] employ boosting for fusion. Each base classifier employed is trained using weighted feature points. These weights are increased if a previous classifier misclassifies. Once trained, their decisions are combined through weighted majority vote [7]. The motion segmentation approach described in [6] is based on a K-nearest-neighbour fusion procedure of labelled spatial and temporal input fields. Fusion is performed using the Iterative Conditional Mode optimisation algorithm of extracted segmentation maps. Three popular optical flow algorithms are fused in [5] using EM reducing the percentage of missing target pixels on a single outdoor dataset by 33%. [1] employ a maximum a posterior (MAP) Bayesian fusion technique. Two separate background models are produced using a Gaussian Mixture Model algorithm and a Brightness and Chromaticity algorithm (as used in this work). When the classifiers agree (pixel is foreground or background) a decision is set accordingly. When classifiers disagree, the conditional probability for the chosen class by each class is calculated. The classifier with the maximum a posterior probability provides the final decision. The main limitation of the approach is that it limited to combination of two classifiers and that the priors are calculated using an exhaustive search method based on the training data to obtain the optimal values giving minimum classification errors.

Some works [9, 10, 14] include pixel-level fusion in conjunction with other modules. For example [14] employ a simple rule-based fusion of separately run flux-tensor based motion detection algorithm and split Gaussian models based background subtraction. Such methods, however, do not isolate the performance of the fusion module and instead concentrate on the end performance of their integrated approaches.

While the above works individually address explicitly [1] or implicitly [5, 6, 9, 10, 14, 16] fusion of outputs insufficient attention has been paid in the literature to quantitative evaluation of different fusion schemes applied to the same data. This work addresses this issue, focussing...
on pixel-level segmentation, applying 4 fusion schemes to the outputs of 3 complementary probabilistic algorithms (Gaussian Mixture Model [12], Colour Mean and Variance [2], and Brightness and Chromacity [13]). The four fusion schemes considered range from the more simplistic (Majority Vote [7], Max Rule [7]) to the more complex information theoretic (Multi Sensor Bayes [8] and Mutual Information [8]).

3. Motion Extraction

Each of the motion classifiers used for this work are probabilistic in nature in that normal distributions are created using $N$ frames of a sequence to create background models based on pixel values. For fusion purposes, the raw probabilistic output of these classifiers is converted to ranked labels Foreground $F$, Highlight $H$, Shadow $S$, and Background $B$ which have assigned probability estimates of 1, 0.5, 0.25 and 0, respectively. A description of the individual algorithms follows.

3.1. Gaussian Mixture Model

The Stauffer and Grimson [12] Gaussian mixture model (GMM) creates a history of a pixel value and is modelled by a mixture of $K$ Gaussians ($K$ usually varies from 3 - 5). The mixture is weighted by the frequency with which each of the Gaussians explains the background. The probability that a pixel has a value $v_i$ at frame $i$ is given by the sum of the weighted probability density for each distribution.

$$P(v_i) = \sum_{j=1}^{K} w_{j,i}N(x, \mu_{j,i}, \Sigma_{j,i})$$  

(1)

where $K$ is the number of distributions, $w_{j,i}$ is a weight associated with the $j^{th}$ Gaussian at frame $i$, $\mu_{j,i}$ is the mean, $\Sigma_{j,i}$ is the standard deviation and $N$ is a Gaussian probability density function. A pixel is classified and assigned a probability $y_i$ of foreground based on the following decision procedure where $G$ is a normalised pixel of the current component and $\hat{S}$ represents a normalised pixel value at frame $i$ after projected onto the components brightness plane $S$.

$$x_i = \begin{cases} 
B: \hat{G}_i < 3.5\sigma & \text{else} \\
S: \hat{S}_i > 0.9\sigma & \text{and} < 1.0\sigma, \text{else} \\
H: \hat{S}_i > 1.0\sigma & \text{and} < 1.2\sigma, \text{else} \\
F: \text{otherwise}
\end{cases}$$  

(2)

3.2. Colour Mean and Variance

The Colour Mean and Variance (CMV) algorithm, based on the algorithm created by Wren [2] builds a statistical background model to represent an independent Gaussian distribution for each normalised colour channel ($R, G, B$) and a Gaussian distribution of the normalised average ($A$) of each pixel colour channel.

$$n(v, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(v-\mu)^2}{2\sigma^2}\right)$$  

(3)

where $v$ is the pixel value of a single channel $R, G, B$, or $A$, $\mu$ is the mean and $\sigma$ is the standard deviation of that channel. A pixel is classified and assigned a probability $y_i$ of foreground based on the following decision procedure where $\hat{R}, \hat{G}, \hat{B}$ and $\hat{A}$ are normalised pixels.

$$y_i = \begin{cases} 
F: \hat{R}_i > 1.4\sigma \text{ or } \hat{G}_i > 1.4\sigma \text{ or } \hat{B}_i > 1.4\sigma \text{ or } \\
A_i > 1.4\sigma, \text{ else} \\
H: \hat{A}_i > 1.0\sigma \text{ and } < 1.2\sigma, \text{ else} \\
S: \hat{A}_i > 0.9\sigma \text{ and } < 1.0\sigma, \text{ else} \\
B: \text{otherwise}
\end{cases}$$  

(4)

3.3. Brightness and Chromacity

This background statistical model, most closely related to the method of [13], distinguishes between the brightness and its chromaticity of any one pixel, over time. Figure 1 represents a graphical representation of the brightness distortion and chromaticity distortion in three dimensional RGB colour space. $E_i$ is the initial (background) colour value for a pixel at frame $i$, and $I_i$ is the current colour value of the image. The line OE from the origin to $E_i$ represents the chromaticity line. Brightness distortion is a scalar value $\alpha$ and scales the point along OE where the orthogonal line from $I_i$ intersects OE. Chromaticity distortion $CD_i$ is the orthogonal distance between the observed colour and the line OE. The values for $\alpha$ and $CD$ are calculated for each of $N$ background frames and the method constructs Gaussian distributions (as in Equation 3) for each normalised colour.
channel (R, G, B) and histograms for the \( \hat{\alpha} \) and \( \hat{CD} \) values. A pixel is classified and assigned a probability \( z_i \) of foreground following two decision procedures below. The first simply classifies a pixel as either foreground or background based on Equation 5.

\[
z_i = \begin{cases} 
F : \hat{R}_i > 2.5\sigma \text{ or } \hat{G}_i > 2.5\sigma \text{ or } \hat{B}_i > 2.5\sigma, \text{ else} \\
B : \text{otherwise}
\end{cases}
\] (5)

If the pixel is classified as background, further evaluation is made using Equation 6.

\[
z_i = \begin{cases} 
F : C_D_i > T_{cd} \text{ else} \\
B : \hat{\alpha}_i < T_{\alpha_1} \text{ and } \hat{\alpha}_i > T_{\alpha_2}, \text{ else} \\
S : \hat{\alpha}_i < 0, \text{ else} \\
H : \text{otherwise}
\end{cases}
\] (6)

Thresholds \( T_{cd}, T_{\alpha_1}, \text{ and } T_{\alpha_2} \) are automatically derived from their histograms such that the cumulative probabilities at that threshold are 0.99 for the \( \hat{CD} \) histogram, and 0.99 using \( T_{\alpha_1} \) and 0.01 using \( T_{\alpha_2} \) for the \( \hat{\alpha} \) histogram.

4. Fusion

The output assigned pixels \( x_i, y_i \) and \( z_i \) (probability of foreground), from the respective BC, CMV and GMM classifiers become the input to the four fusion schemes.

4.1. Probabilistic

4.1.1 Multi Sensor Bayes

Given Bayes Theorem for events \( a, b \) may be expressed as

\[
P(a|b) = \frac{P(b|a)P(a)}{P(b)}
\] (7)

then the probability of a pixel in a Bayesian fused motion mask \( m \) at frame \( i \) being foreground \( FG \) given our classifier pixel output \( x, y, z \) (from the CMV, GMM and BC classifiers respectively) is

\[
P(m_{iFG}|x_i, y_i, z_i) = \frac{P(x_i, y_i, z_i|m_{iFG})P(m_{iFG})}{P(x_i, y_i, z_i)}
\] (8)

If we assume independence as the output of one background or foreground model does not affect another then

\[
P(m_{iFG}|x_i, y_i, z_i) = \frac{P(x_i|m_{iFG})P(y_i|m_{iFG})P(z_i|m_{iFG})P(m_{iFG})}{P(x_i, y_i, z_i)}
\] (9)

A training set of \( N \) frames is used to create a prior of all classified background or foreground pixels \( M \) using corresponding frames of binary mask ground truth. \( M \) is used to create the initial history of \( X|M_{FG}, Y|M_{FG} \) and \( Z|M_{FG} \) using the same \( N \) training set frames.

4.1.2 Mutual Information

Mutual information is a well established fusion technique used for medical image registration of both the same and differing modalities [3, 11]. The benefits of its use stem from its insensitivity to changes in lighting conditions and its ability to address a wide range of non-linear image transformations. In information theory the entropy of a discrete random variable \( X \) is the measure of the amount of uncertainty associated with the value of \( X \). If \( p \) represents a probability mass function of \( X \) then Shannon entropy, denoted by \( H \), is described in terms of a discrete set of probabilities with the following formula

\[
H(X) = - \sum p(x) \log p(x).
\] (10)

Mutual information \( I \) measures the amount of information that can be obtained about one random variable by observing another. Mutual information can be expressed in terms of entropies as

\[
I(X; Y) = H(X) - H(X|Y)
\]

where \( H(X) \) and \( H(Y) \) are the marginal entropies, \( H(X, Y) \) is the joint entropy, \( H(X|Y) \) and \( H(Y|X) \) are the conditional entropies and \( H(Y|X) \) is a measure of what \( Y \) does not say about \( X \). \( I(X;Y) \) is non-negative.

Bivariate and trivariate mutual information defined in terms of entropies is shown in Figure 2. Here it is used as a measure of the information or interaction between pixels output by the three algorithms at any given time. To this end, \( CMI \) (Combined Mutual Information) is defined as a linear combination of trivariate mutual information for all three classifiers and bivariate mutual information for each pair of classifiers such that none of the constituent entropies are counted twice. Bivariate mutual informations are \( I(X; Y), I(X; Z) \) and \( I(Y; Z) \) (Figure 2 a, b, and c respectively). The quantity \( I(X; Y) - I(X; Y|Z) \) in Figure 2 d. may be defined as

\[
I(X; Y) - I(X; Y|Z) = I(X; Z) - I(X; Z|Y) = I(Y; Z) - I(Y; Z|X).
\] (12)
the output of three motion classifiers $CMI$ (ensuring no
overlapping entropies) is expressed as

$$CMI = H(X, Y) + H(X, Z) + H(Y, Z)$$

$-2H(X, Y, Z).$  \hfill (13)

With reference to Equation 10 $CMI$ may be rewritten as

$$CMI = -\sum_{x,y} p(x, y) \log p(x, y)$$

$$-\sum_{y,z} p(y, z) \log p(y, z)$$

$$-\sum_{x,z} p(x, z) \log p(x, z)$$

$$+2 \sum_{x,y,z} p(x, y, z) \log p(x, y, z)$$ \hfill (14)

and yields an expected value over all possible instances of
$X, Y$ and $Z$. It is therefore possible to define a quantity
$pCMI$ [4] for each point based on the point wise con­stituents of $CMI$, which provides the calculation of the
mutual information of all three classifiers BC, CMV and
GMM, for each of the respective probabilities assigned to
pixel outputs $x, y$ and $z$ in each frame $i$:

$$pCMI_i = -p(x_i, y_i) \log p(x_i, y_i)$$

$$-p(y_i, z_i) \log p(y_i, z_i)$$

$$-p(x_i, z_i) \log p(x_i, z_i)$$

$$+2p(x_i, y_i, z_i) \log p(x_i, y_i, z_i)$$ \hfill (15)

4.2. Majority Vote

A pixel-based majority vote approach is implemented as
the third fusion scheme. For each pixel $(x_i, y_i$ and $z_i)$
of each frame a count $F_{count}$ is made of how many times that
pixel has been classified as $F$ by their respective classifier
BC, CMV and GMM. A resulting pixel $mv_i$ is fused with the
following procedure:

$$mv_i = \begin{cases} F : F_{count} >= 2 & \text{else} \\ B : \text{otherwise} \end{cases}$$ \hfill (16)

4.3. Max Rule

The final fusion scheme in this work classifies a result­ing
$mri$ pixel simply by assigning pixel $mr_i$ with the same
probability and label of the pixel $x_i, y_i$ or $z_i$ which has
been classified by BC, CMV and GMM respectively with
the greatest probability of being Foreground.

$$mr_i = \max(x_i, y_i, z_i)$$ \hfill (17)

5. Experimental results

5.1. Datasets

The choice of datasets for this work were based on the
availability of motion segmentation ground truth and rep­resentation of the diversity of outdoor surveillance scenes.
Hence, the categories Baseline, Bad Weather, Camera Jitter,
Dynamic Background, Shadow and Thermal challenges
from the changedetection.net website [15] were employed.

5.2. Metrics

Results are presented by reporting the performance of
each fusion method, as well as each individual method, for
each video category with respect to 7 different performance
metrics. Let $TP$ be the number of true positives, $TN$ be the
number of true negatives, $FN$ the number of false negatives,
and $FP$ the number of false positives. For each fusion ap­proach and each video, the following 7 metrics are com­puted:

1. Recall (Re): $TP / (TP + FN)$
2. Specificity (Sp): $TN/(TN+FP)$
3. False Positive Rate (FPR): $FP/(FP+TN)$
4. False Negative Rate (FNR): $FN/(TN+FP)$
5. Percentage of Wrong Classifications (PWC):
   \[ \frac{100(FN+FP)}{TP+FN+FP+TN} \]

6. Precision (Pr):
   \[ \frac{TP}{TP+FP} \]

7. F-measure (or F1 score):
   \[ 2 \frac{Pr \cdot Re}{Pr + Re} \]

5.3. Evaluation

From Table 1 and Figure 3 it is clear that except for the MAX fusion technique, all other 3 fusion techniques outperform each individual input method, with Majority Vote ranking overall highest across all categories (Table 2). This reinforces the belief that fusion based methods combine the strengths of each individual algorithm. As for the MAX fusion technique, even though it produces the best performance for Recall and False Negative Rate metrics as shown in Figure 3, it is ranked bottom for Specificity and only second last for F-Measure. This proves that simply taking the largest probability value is not a good strategy (see Figure 4 for an example), since more foreground pixels are gained at the expense of more false classifications and noise. Referring to Figure 3 all three individual motion classifiers and four fusion techniques produced a minimum false positive rate FPR and mostly maximum Specificity score. Out of the single motion classifier techniques GMM proved strongest in performance, where for both the Recall and F-Measure metric it was only outperformed by the MAX fusion technique. Referring to the left column of Figure 4 the motion masks clearly show the difficulty that individual algorithms have in dealing with challenging environments such as the waterfall scene. It can be seen from Figure 4 that the fusion of the individual classifiers BC, CMV and GMM using both Bayesian fusion Bayes and Majority Vote fusion MV visually produces better results.

6. Conclusions and future work

The fusion of the output of motion classifiers has been demonstrated to be a realistic technique to improve the quality of the resulting motion masks generated across a range of datasets. Whilst the Bayesian fusion technique Bayes ranks top for three of the six categories, the Majority Vote MV technique may be the best choice of fusion for universal use, particularly where the challenges for motion classification within a video category is unknown. This result confirms the findings reported on the changedetection.net website where majority vote is the top scoring classifier. Future work aims to investigate the loss of overall accuracy that occurs for certain categories due to the suppression of the results of one or more accurate algorithms by consistent errors of the other two algorithms, and to consider a larger set of dataset categories and algorithms.

References

Figure 4. Representative category results. Left column: Dynamic Background; top-to-bottom: original image, ground truth, Bayes, MV, GMM, BC, MI, MAX, CMV. Right column: Thermal; top-to-bottom: original image, ground truth, MY, MI, BC, Bayes, MAX, GMM, CMV (see Table 2 for category ranks).


