Change Detection in Crowded Underwater Scenes
Via an Extended Gaussian Switch Model Combined with a Flux Tensor Pre-segmentation

Martin Radolko, Fahimeh Farhadifard and Uwe von Lukas
Institute for Computer Science, University Rostock, Rostock, Germany
Fraunhofer Institute for Computer Fraphics Research IGD, Rostock, Germany
{*Martin.Radolko, Fahimeh.Farhadifard*}*@uni-rostock.de, Uwe.Freiherr.von.lukas@igd-r.fraunhofer.de*

Keywords: Change Detection, Background Subtraction, Video Segmentation, Video Segregation, Underwater Segmentation, Flux Tensor.

Abstract: In this paper a new approach for change detection in videos of crowded scenes is proposed with the extended Gaussian Switch Model in combination with a Flux Tensor pre-segmentation. The extended Gaussian Switch Model enhances the previous method by combining it with the idea of the Mixture of Gaussian approach and an intelligent update scheme which made it possible to create more accurate background models even for difficult scenes. Furthermore, a foreground model was integrated and could deliver valuable information in the segmentation process. To deal with very crowded areas in the scene – where the background is not visible most of the time – we use the Flux Tensor to create a first coarse segmentation of the current frame and only update areas that are almost motionless and therefore with high certainty should be classified as background. To ensure the spatial coherence of the final segmentations, the N^2Cut approach is added as a spatial model after the background subtraction step. The evaluation was done on an underwater change detection datasets and showed significant improvements over previous methods, especially in the crowded scenes.

1 INTRODUCTION

The detection of objects in videos has already a long history in computer vision but still is a very relevant topic today due to new developments such as self driving cars or robot aided production which demand a detection in real time and with high precision. In this paper, we address the specific topic of the segregation of a video into two parts, the static background and the moving foreground. This is an important first step in a computer vision pipeline since moving objects are almost always the most interesting part of a scene. For example, if a car or robot has to avoid collisions, then the objects that are moving pose the highest threat and knowledge about their exact position and direction of movement is mandatory.[1]

To detect these moving objects we assume a static camera, so that stationary objects also appear stationary in the video. This makes it possible to create a model of the static background of the scene, e.g. with statistical methods, and every object that does not fit the model is therefore labeled as a moving object. In recent years many of these background modeling and subtraction algorithms have been proposed, but as the tasks and applications of these methods are as plentiful as the suggested algorithms there is still a lot of research to be done.

In this paper, we focus on crowded scenes which pose a particularly difficult task for background subtraction algorithms since the permanent exposure to foreground objects often leads to an adaption of the background model to these foreground objects, especially when they are all similar in color like the fishes in a swarm. To cope with this we introduce pre-segmentations created with a Flux Tensor-based optical flow which are used to exclude parts of the current frame from the updating process of the background model. These parts are very likely to be foreground since they are in motion and therefore excluding them limits the background modeling to the background parts of the scene.

Furthermore, we enhance the Gaussian Switch Model approach proposed in (Radolko and Gutzeit, 2015) with the Mixture of Gaussian idea, a fore-

[1]This research has been supported by the German Federal State of Mecklenburg-Western Pomerania and the European Social Fund under grant ESF/IV-BMB35-0006/12]

Radolko M., Farhadifard F. and Lukas U.
Change Detection in Crowded Underwater Scenes - Via an Extended Gaussian Switch Model Combined with a Flux Tensor Pre-segmentation.
DOI: 10.5220/0006258504050415
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 405-415
ISBN: 978-989-758-225-7

ground model and an intelligent updating scheme to make it overall more robust for difficult scenarios. The foreground model proved to be particularly useful in the scenes with fish swarms because the difference between the different foreground objects was minor and thereby the time for the model to adapt to a new object was negligible. Lastly, since the approach so far is solely pixel-based, a spatial component was added to make the segmentations coincide with the edges in the frame and better conform to the smoothness of natural images.

2 STATE OF THE ART

Background modeling and subtraction has been used in computer vision for a long time already. The first approaches date back to the beginning of the 90ths (Ridder et al., 1995; Shelley and Seed, 1993) and commercial applications followed soon. An example is the Patent (Gardos and Monaco, 1999), where background subtraction is used for video compression.

The most frequently used approaches in recent years have been statistical methods that use gaussians to model each pixel of the background. It started with the Single Gaussian approach (Wren et al., 1997) where one Gaussian distribution is used to describe one pixel value. They are usually updated with a running gaussian:

$$m_{t+1} = \alpha m_t + (1-\alpha)p. \quad (1)$$

Here m_t is the mean of the gaussian at the time step t,p is the pixel value taken from the current frame and $\alpha \in (0,1)$ is the update rate.

However, this simple method is not sufficient to model difficult scenes – e.g. permanent exposure to many foreground objects or slightly moving background objects – and therefore in (Stauffer and Grimson, 1999) an algorithm was proposed which does not use only one gaussian but a mixture of several gaussians. This proved to be a very effective way of modeling the background and is henceforth used with great success in combination with other methods. Together with a Markov Random Field the Mixture of Gaussian (MoG) is used in (Schindler and Wang, 2006) and can generate great results on the *Wallflower* dataset. In conjunction with an optical flow, the Flux Tensor, it is used in (Wang et al., 2014) and achieves state of the art results on the *changedetection.net* dataset.

The Sample Consensus methods take another approach by keeping a set of samples for each pixel position instead of modeling the color of that pixel directly in a probability distribution. The *ViBe* algorithm (Barnich and Droogenbroeck, 2011) is one example for this class, it updates the samples for each pixel randomly so that even old values can have an influence on the current segmentation (although with a decreasing probability). Furthermore, the updating process diverges spatially so that an update of one sample can influence the neighbouring samples which makes the model spatially coherent to some degree. The segmentation itself is done by counting the number of values that agree with the current value, which means that they are closer to the value than a specific threshold. If enough samples agree with the current pixel, it is assumed to be background.

The approach of (St-Charles et al., 2015) is similar to that but does not store the pixel values directly but rather in a feature vector based on Local Binary Similarity Patterns (LBSP) that describes the pixel and its neighbourhood. Furthermore, they used a sophisticated scheme to adapt their parameters to the current situation based on a regional classification. Their segmentation quality and runtime can compete with state of the art approaches .

A non-parametric algorithm is proposed in (Zivkovic and Heijden, 2006) by using a k-Nearest Neighbors approach and a good implementation of this is freely available in the OpenCV library. In (Marghes et al., 2012; Hu et al., 2011) the Principal Component Analysis (PCA) is used to extract the background of a video and the non-negative matrix factorization was used similarly in (Bucak et al., 2007). However, these subspace approaches can generally not achieve results equivalent to the aforementioned methods and are also often computationally very expensive. A background model based on a Wiener Filter in conjunction with a regional approach which smooths the segmentation and adapts it to the edges of the current frame was introduced in (Toyama et al., 1999). Also, there is mechanism that monitors the whole frame to find global changes, e.g. a light that is switched off and makes the whole scene appear dark.

There are also approaches which automatically combine whole segmentations of various methods in a way that the output is better than each individual input. Examples are (Mignotte, 2010), which used a Bayesian Model and a Rand Estimator to fuse different segmentations, or (Warfield et al., 2004) which applies Markov Random Fields to fuse segmentations of medical images. A quite current approach is (Bianco et al., 2015) which uses the large database of different segmentations of the *changedetection.net* dataset and combines the best performing of them. The fusion process itself is not done by a Bayesian Model, like

in the other cases, but with a genetic algorithm. The genetic algorithm has the segmentations and a set of functions it can apply on them and tries to find the best combination. These functions are e.g. morphological erosion or dilation, logical *AND* or *OR* operations or a majority vote on the binary segmentations.

In this way they can improve the already very good results of the top algorithms. However, to run their genetic algorithm groundtruth data is necessary and therefore they use one video of each category (and the corresponding groundtruth data) to find the best combination of segmentations and functions. They can achieve better results than all the existing algorithms with this approach but the need of several already good segmentation results and known groundtruth data for the training phase makes this approach impractical.

3 PROPOSED APPROACH

The proposed method consists of three steps. The first step is explained in the sections one and two where we describe the Gaussian Switch Model (GSM) and introduce our extension of it. In the next section we derive segmentations based on the Flux Tensor and use them to improve the background modeling process of the extended GSM. The last step in section four is a spatial approach which adapts the segmented objects of the background subtraction to the edges in the image by using a NCut based approach.

3.1 Gaussian Switch Model

The GSM was introduced in (Radolko and Gutzeit, 2015) and models the background of the scene with two distinct gaussian models for each pixel in the video. Of these two models, one is updated conservatively (only parts classified as background are updated) and one is updated blindly (the whole image is updated) which allows the method to benefit from the advantages of both strategies.

The conservative strategy has the problem that rapid changes of the background will not get incorporated into the model, an example of this would be a car that parks and therefore, after some time, should become a part of the background model. The blind update strategy, on the other hand, has the problem that the foreground objects get included into the background model as well and especially in scenes with a constant presences of foreground objects this can lead to a corrupted background model.

The GSM now has models with both of these updating strategies and normally the one with the conservative updating strategy is used for the background subtraction because it creates a clearer and more accurate background model in most situations. However, scenes in which the conservative update model fails can be detected by a comparison of both models and if such a situation is detected, then the model is switched from the conservative updated one to the model which was blindly updated. A depiction of the effect of these different strategies can be seen in Figure 1.

Another omnipresent problem in background subtraction is shadow, to make the approach more robust against changes of the lightning a special color space is used. The conversion is done with the following equation

$$\begin{aligned} I &= R+B+G, \\ C_1 &= R/I, \\ C_2 &= B/I. \end{aligned} \tag{2}$$

Afterwards the intensity I is scaled with the factor $\frac{1}{3 \cdot 255}$ so that all values are in the range $[0,1]$. The values C_1 and C_2 only contain color information and should not change if a shadow appears or the lightning conditions of the scene change.

3.2 Extension of the GSM

We propose an extension of the GSM background modeling method by superimposing it with the Mixture of Gaussian idea. This makes the whole approach more complex and slower but also can increase the accuracy further, especially in difficult situations like the underwater scenes we use for evaluation later.

Instead of using two Single Gaussian models we apply two Mixture of Gaussian models and update one of them conservatively and one blindly. Also, we added a foreground model with a high adaption rate to quickly adapt to different moving objects in the scene. We chose a simple single gaussian model for this because it should not model different foreground objects at the same time but only the most recent one.

Each Mixture of Gaussian (MoG) consists of a variable number of gaussians (we used five) and each of them is described by three values: mean m, variance v and weight w. The mean and variance describe the shape of the probability distribution and the weight is a measure of how much data supports this gaussian. To be considered as a part of the background model a minimum weight is necessary, otherwise the gaussian is assumed to belong to a foreground object which only appeared shortly in the video. We define the minimum weight as a percentage of the sum over all weights of a MoG and set the percentage to $1/\#\text{gaussians}$, so one fifth in our case.

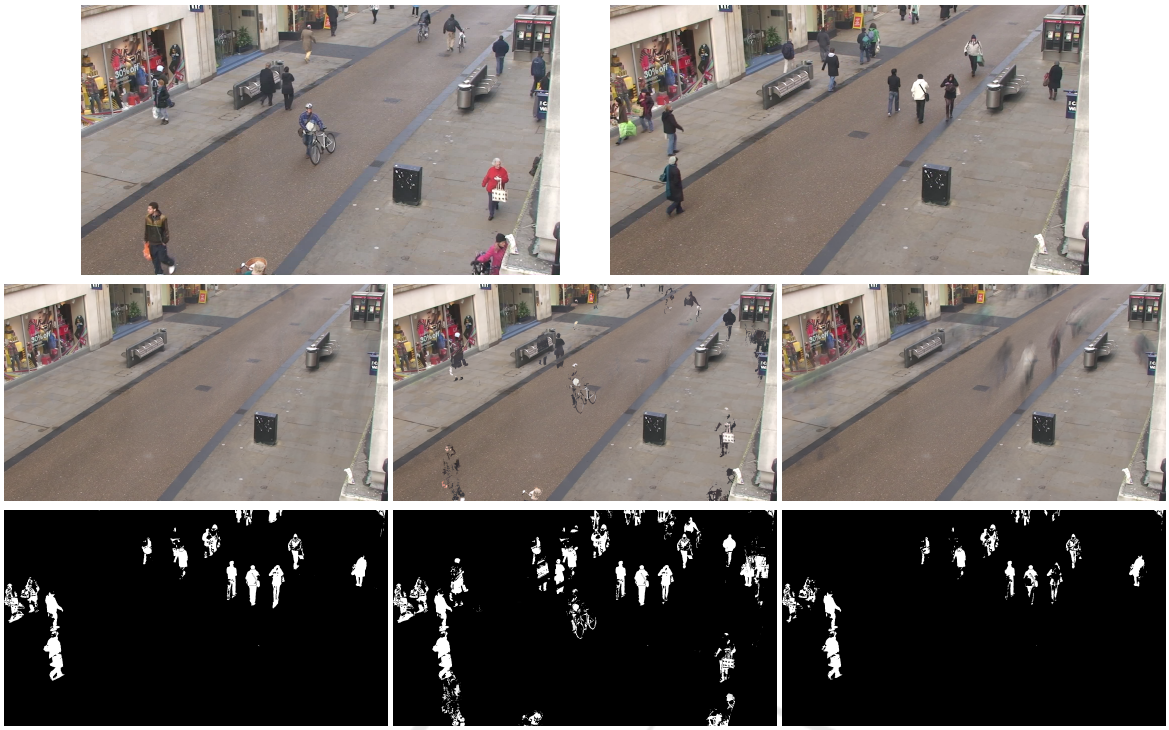

Figure 1: Comparison of different update schemes for the background modeling. In the top row are the first and 2000th frame of the Town Center video from (Benfold and Reid, 2011). In the next row are three background models for the 2000th frame of the video created with the same parameters but different updating mechanisms: the first was created with the GSM approach, the second with a purely conservative updating scheme and the last one with a blind update for every frame. The conservative model still has many artefacts from the first frame as they were always marked as foreground and therefore never updated. The blind update creates a model corrupted with foreground information from the recent frames and only the combination of them in the GSM could create an accurate background model. The last row shows the corresponding segmentation for every model.

The MoGs are updated by first searching for the gaussian that matches the current data the best and then applying the standard running gaussian update on them. For a pixel x with pixel value p_x and update rate α the equations would be the following

$$\begin{aligned} v_x &= \alpha \cdot v_x + (1-\alpha) \cdot (m_x - p_x)^2, \\ m_x &= \alpha \cdot m_x + (1-\alpha) \cdot p_x, \\ w_x &= w_x + 1. \end{aligned} \tag{3}$$

The α value is specified dynamically according to the weight value of the gaussian in the following way

$$\alpha = \frac{1}{w_x} \tag{4}$$

but it is capped at 0.5. Furthermore, to prevent an overflow of the weight value and limit the impact of old values on the model, there is a decay of all weight values in the MoG.

Together, this ensures that gaussians which until now only got very few datapoints to back them up or only old datapoints which are not reliable anymore adapt quickly to new values. At the same time, gaussians which were updated frequently (and therefore have a high weight) will get a small α and are not strongly effected by single outliers. Consequently, the decay factor has a strong impact on the update rate, especially in longer videos, and is therefore the most important parameter. Empirically we choose it to be 0.995 in our experiments, that means the sum of all weights in a MoG will tend to 200 for longer periods.

If no matching gaussian could be found in the existing MoG model a new gaussian will be created with the values $m_x = p_x$, $v_x = 0.01$ and $w_x = 1$. Should there already exist the maximum number of gaussians that are allowed, the gaussian with the lowest weight will be deleted and replaced with the new one.

The foreground model is also updated as a running gaussian but with a fixed α_F value as there is no weight value in the Single Gaussian model. Also, the update rate should be higher than in the background models so that it can adapt quickly to new foreground objects. We set it to $\alpha_F = 0.64$ for our experiments. Nonetheless, before the updating process of the model

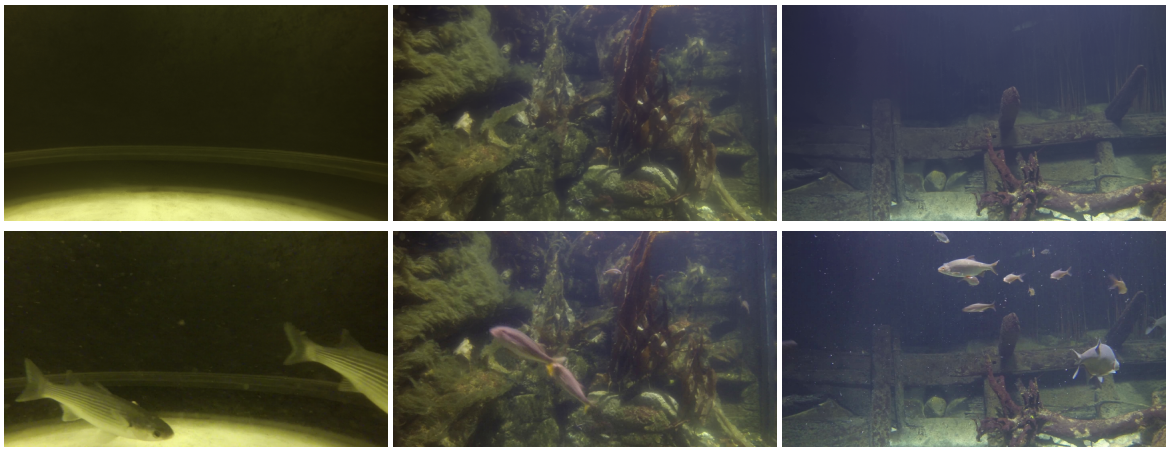

Figure 2: The top row depicts three background models created with the extended GSM and below that are the corresponding original frames from the video. The background models are visualized by taking the gaussian with the highest weight of the conservatively updated MoG and displaying the mean of it.

starts, the segmentation has to be done with the existing model and based on this result the different models will get modified accordingly.

The blindly updated MoG is updated every time regardless of the segmentation result. The conservative MoG only gets updated when a pixel was classified as background and the foreground model obviously only when the pixel was marked as foreground. The segmentation itself is created by comparing the current frame with the two MoGs. However, only the gaussians that have a weight that exceeds the minimum weight (one fifth of the overall weight) are considered part of the background model. If for any of these gaussians the inequality

$$\exp\left(-\left\|\frac{1}{\beta}\cdot\frac{\bar{p}_x-\bar{m}_x}{\bar{v}_x}\right\|_2^2\right)>0.5 \qquad (5)$$

is true, the pixel value and the MoG are classified as a match. The vectors $\bar{p}_x$, $\bar{m}_x$ and $\bar{v}_x$ contain the values of the three channels of the pixel x and the operations between them are all elementwise. The variance as a divisor makes the thresholding process adaptive, so that it is less sensitive if the video contains few noise and vice versa. The value β in the inequality is a parameter controlling the general sensitivity of the approach and we set it with 0.5 quite low since the foreground objects in our data are often quite similar to the background and therefore a high sensitivity is necessary.

If the pixel matches with gaussians in both MoGs, it will be classified as background. However, if it only matches with one of the MoGs the foreground model is taken as a tiebreaker. The foreground model is compared to the pixel value according to the inequality (5) and if it matches the pixel it is marked as foreground, otherwise as background.

Similar to the original GSM algorithm, there is a switching between the conservatively updated MoG and the blindly updated MoG to compensate for the weaknesses of conservative updating scheme. Such a switch should occur when there is something in the scene which is static and constantly classified as foreground, because then, with a high probability, an error in the background modeling happened and should be corrected.

To detect such an error the first condition is that the blindly updated MoG and the foreground model are similar since this indicates that this pixel has been mainly classified as foreground in the recent past. The models are considered similar if

$$m^{BG,k}-m^{FG}<\frac{v^{FG}}{2} \qquad (6)$$

holds for all three channels of a pixel. Here $m^{BG,k}$ is the mean of the k-th gaussian of the conservatively updated MoG and it is sufficient if the inequality is true for one of the gaussians of a pixel. This similarity could also occur when there appear many foreground objects in a short period of time. To filter these events out the variance can be used since foreground objects usually generate higher variations in the image due to their movement. Hence the second conditions is a small variance and the threshold is set to the median of all variances of the completely updated MoG. If both of these conditions are fullfilled (inequality 6 and small variance) an error in the conservatively updated MoG is very probable and therefore the blindly updated MoG is used in these cases.

Lastly, it can occur that two gaussians in one MoG get very similar over time. These gaussians then should be unified as they are modeling the same object. The similarity is checked with

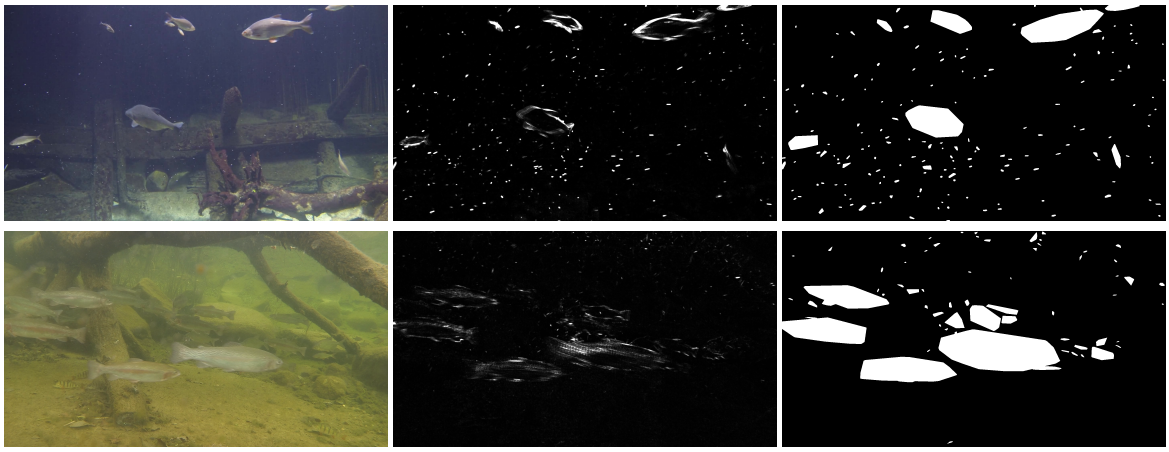

Figure 3: The Flux Tensor on two examples with fishes as moving objects. The images in the middle show the result of the actual Flux Tensor, higher intensities depict higher movement. On the right side is the segmentation after clustering and building a convex hull around the foreground clusters. The noise, especially in the upper example, is due to the Marine Snow which are small floating particles.

$$\|\tilde{m}^{G1} - \tilde{m}^{G2}\|_2^2 < \min(\|\tilde{v}^{G1}\|_2^2, \|\tilde{v}^{G2}\|_2^2) \tag{7}$$

and if the inequality holds, the old gaussians are deleted and a new gaussian is created with the following values

$$\begin{aligned} m^{new} &= \frac{w^{G1}m^{G1} + w^{G2}m^{G2}}{w^{G1} + w^{G2}}, \\ v^{new} &= \frac{w^{G1}v^{G1} + w^{G2}v^{G2}}{w^{G1} + w^{G2}}, \\ w^{new} &= w^{G1} + w^{G2}. \end{aligned} \tag{8}$$

Altogether, this extension of the standard GSM leads to a robust and accurate model building process since now several different objects can be represented by the model at the same time and the update rate adapts itself automatically based on the confidence the model has in the data. Three examples of modeled backgrounds can be seen in Figure 2.

3.3 Flux Tensor as a Pre-segmentation

Two dimensional structure tensors have been widely used for edge and corner detection in images, e.g. in (Nath and Palaniappan, 2005). They use the information of derivates of the images and are applied as filters on the image which makes them computationally very efficient. Motion information can be recovered in a similar way, but then there has to be a three dimensional tensor which is applied on an image volume of a video.

For the location $p = (x, y, t)$ in an image volume the optical flow $v(p) = [v_x, v_y, v_t]$ is usually computed with the formula

$$\frac{\partial I(p)}{\partial x}v_x + \frac{\partial I(p)}{\partial y}v_y + \frac{\partial I(p)}{\partial t}v_t = 0 \tag{9}$$

which leads to an eigenvalue problem that is costly to solve. To extract the valuable motion information without solving the eigenvalue problem the flux tensor was proposed in (Bunyak et al., 2007) and is defined by

$$\begin{aligned} &\int_{p\in\Omega} \left(\frac{\partial^2 I(p)}{\partial x \partial t}\right)^2 + \left(\frac{\partial^2 I(p)}{\partial y \partial t}\right)^2 + \left(\frac{\partial^2 I(p)}{\partial t \partial t}\right)^2 dz \\ &= \int_{p\in\Omega} \|\frac{\partial}{\partial t}\nabla I(p)\|^2 dz, \end{aligned} \tag{10}$$

for the pixel p and a small area Ω around it. By computing the Flux Tensor one value per pixel is obtained which represents the magnitude of motion in that area (but not the direction of the movement) and this can be thresholded to get a binary segmentation.

However, when objects are uniform, the Flux Tensor did have difficulties segmenting the interior of the objects and often only detected the edges as moving. To cope with this behaviour we use a density-based spatial clustering after the thresholding and then create a convex hull around these clusters of foreground detections. This method can detect moving objects very reliable but the created segmentation does not reflect the actual shape of the objects very well. Two examples of both steps of the algorithm can be seen in Figure 3.

Although these segmentations are in general not as accurate as those derived from a background subtraction approach they have the advantage to be available without a learning phase which can be very useful. The extended GSM has a very elaborated learning

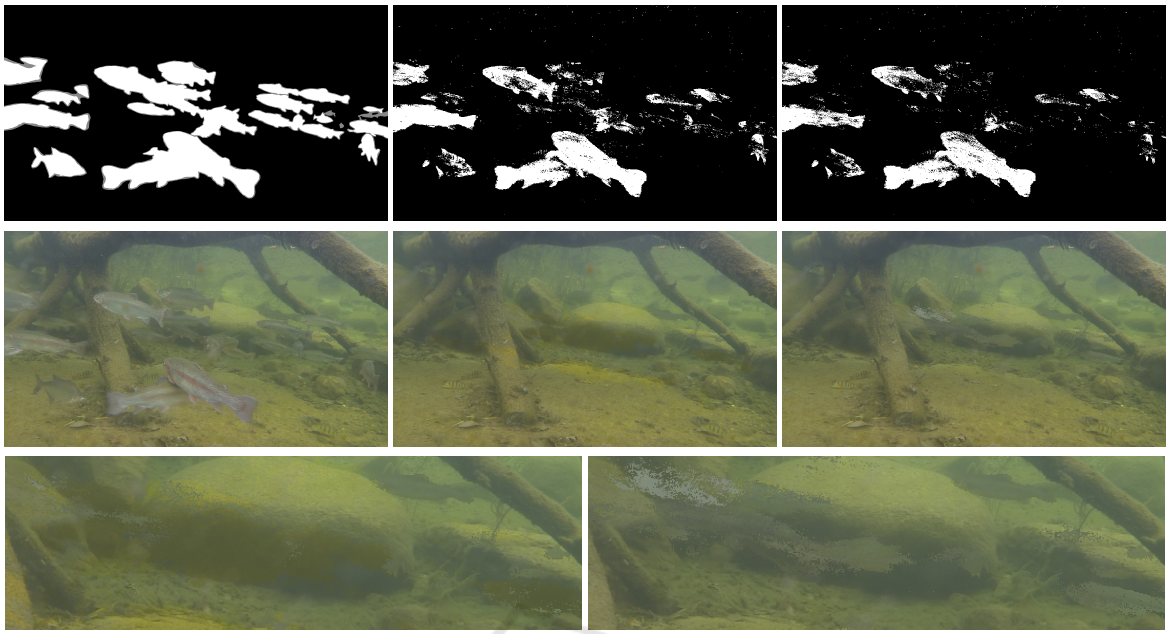

Figure 4: Effect of the Flux Tensor pre-segmentations on the background modeling. In the top row are from left to right the ground truth image, the segmentation of the extended GSM method with pre-segmentation and without pre-segmentation. Below that are the orignal frame and the corresponding visualizations of the two background models. The last row shows a close up of the background models in an area where many fishes were passing by. The model created with pre-segmentations (left) has less artefacts of fishes and is also not as blurry.

algorithm but there are still problems in very crowded scenes. This is caused by an inherent problem in the modeling of the background: it assumes that the background objects are visible the majority of the time and will therefore adapt to the objects that appear the most.

This is true in almost all of the background subtraction scenarios and works very well. However, in some of the underwater scenes that we address here, there is a fish swarm in a certain area and most of the time fishes are visible there and not the real background. Therefore, the background model would adopt to the color of the fishes and not to that of the background. To solve this problem we use the Flux Tensor segmentations as a mask for the updating of the background model. Thereby, areas with high movement are not updated since it would only train the model with information about foreground objects.

The principle is similar to that of the conservative updating scheme which excluded pixels that are classified as foreground from the updating. However, this does only work if the background model is already accurate and a good segmentation can be provided. In a scene that is constantly crowded no good background model can be created and therefore the conservative updating scheme fails. Here the pre-segmentations can help since they do not need any model and help creating a proper background model in the first place. An example of this effect can be seen in Figure 4. The visualization of the background model is created by taking the gaussian with the highest weight of the conservatively updated MoG and displaying the mean of it.

3.4 N^{2}Cut

Until now, the whole approach is completely pixel-wise and only uses the temporal changes to detect foreground objects. However, natural images have spatial properties that can be used to further improve the derived segmentations, e.g. a certain degree of smoothness is always present and edges in the segmentation should be aligned to edges of the frame since they often represent borders of objects.

To this end, we use the N^{2}Cut from (Radolko et al., 2015) here. It is a GraphCut based approach with a special energy function derived from NCut. The NCut is defined as

$$\begin{aligned} NCut(A,B) &= \frac{Cut(A,B)}{Assoc(A)} + \frac{Cut(A,B)}{Assoc(B)}, \\ Assoc(A) &= \sum_{i \in A, j \in A \cup B} w_{ij}, \\ Cut(A,B) &= \sum_{i \in A, j \in B} w_{ij}, \end{aligned} \tag{11}$$

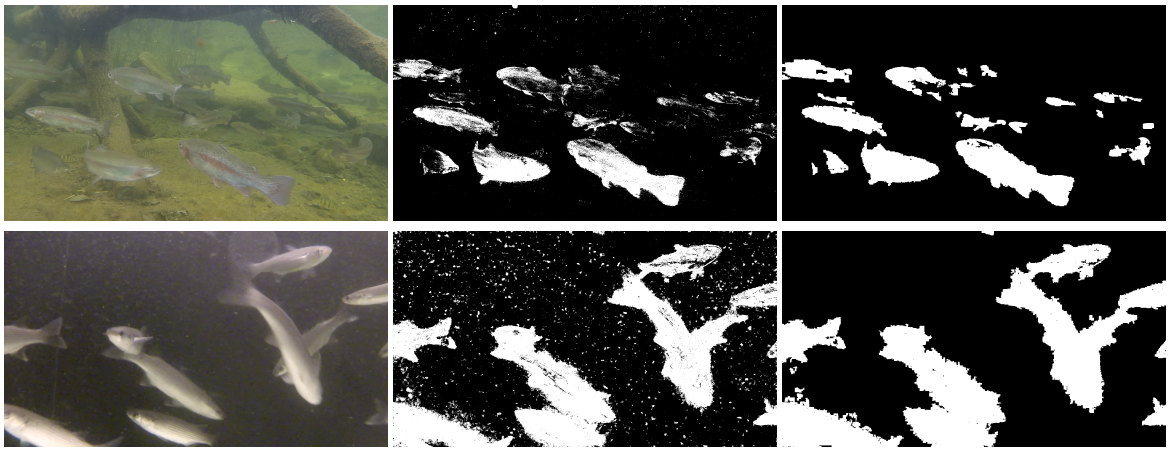

Figure 5: Example of the effect of the N^{2}Cut method. On the left are the original images, in the middle the segmentations after background subtraction and on the right is the result after applying the N^{2}Cut.

where A and B are the sets of foreground and background pixels and w_{ij} is a weight function. It is defined as a sum over the three channels of the pixels i and j by

$$w_{ij} = |r_i - r_j| + |g_i - g_j| + |b_i - b_j|, \qquad (12)$$

if the pixels are neighbors and is 0 otherwise. Based on this, the N^{2}Cut is defined as

$$\begin{aligned} N^2Cut(A,B) &= \frac{Cut(A,B)}{nAssoc(A)} + \frac{Cut(A,B)}{nAssoc(B)}, \\ nAssoc(A) &= \frac{Assoc(A)+1}{\sum_{i\in A, j\in A\cup B, \exists e_{ij}} 1 + 1}. \end{aligned} \qquad (13)$$

In this new energy function the *Cut* and *Assoc* values are normalized by number of elements that contribute to them. Thereby, it still favors segmentations that are aligned with edges in the image, similar to the NCut, but also is free of any bias for a certain amount of background or foreground in the segmentation whereas the NCut tends to segmentations with an equal amount of fore- and background. This is an important feature for video segmentation as there are often times when no foreground objects at all are present in the scene.

This energy function will now be minimized over the already existing segmentation derived from the background subtraction. To this end a local optimization is applied by changing the classification of single pixels which are located at the border between foreground and background areas. The new N^{2}cut value, after changing only pixel d from set A (foreground) to B (background), can be computed very efficiently with just a few additions and subtractions by using the following formulas

$$\begin{aligned} &Cut(A \setminus \{d\}, B \cup \{d\}) = \\ &Cut(A,B) + \sum_{i\in A \,\wedge\, i\in N(d)} w_{id} - \sum_{j\in B \,\wedge\, j\in N(d)} w_{jd}, \end{aligned} \qquad (14)$$

$$\begin{aligned} &Assoc(A \setminus \{d\}) = \\ &Assoc(A) - \sum_{i\in B \,\wedge\, i\in N(d)} w_{id}, \end{aligned} \qquad (15)$$

$$\begin{aligned} &Assoc(B \cup \{d\}) = \\ &Assoc(B) + \sum_{i\in A \,\wedge\, i\in N(d)} w_{id}. \end{aligned} \qquad (16)$$

Here $N(d)$ is the four connected neighborhood region of d. Thereby, the N^{2}cut value and the segmentation can be gradually improved without the high computational cost of the global optimization of a cut value over a whole image. To increase the range/effect of the minimization we apply it over several scales of the image, starting with the smallest size and using the result from there as a starting segmentation for the next scale. Overall, this proved to be an efficient way to smooth the segmentation derived from the background subtraction and align it to the edges of objects in the frame. An example is depicted in Figure 5.

4 RESULTS

For the evaluation we took the dataset and numbers presented in (Radolko et al., 2016) [2]. It is the only underwater change detection dataset so far and includes five videos of different scenes with fishes as moving foreground objects. For each video the first 1000

[2]dataset available at: underwaterchangedetection.eu

Table 1: The results of our and four different background subtraction methods on the underwater change detection dataset. The first one is the original GSM algorithm, the next two are MoG approaches and the last one is a background modeling method based on K-nearest neighbours. The amount of foreground (TP+FN) is not constant because of the small uncertainty area in the dataset, for details see (Radolko et al., 2016).

Algorithm	True Negative	True Positive	False Negative	False Positive	F1-Score
(Radolko and Gutzeit, 2015)	892,599,998	84,349,046	44,709,358	17,215,198	0.7314
(Zivkovic, 2004)	897,659,412	76,555,440	52,934,753	11,723,995	0.7030
(Kaewtrakulpong and Bowden, 2002)	912,653,089	36,184,999	86,673,758	1,288,154	0.4513
(Zivkovic and Heijden, 2006)	887,967,097	93,919,051	36,656,258	20,331,194	0.7672
proposed	872,360,637	116,926,034	30,167,642	17,345,687	0.8781

Table 2: The F1-Scores of the algorithms from Table 1 for each video of the dataset separately.

Algorithm \ Video	Fish Swarm	Marine Snow	small Aquaculture	Caustics	two Fishes
(Radolko and Gutzeit, 2015)	0.5691	0.8361	0.7734	0.5499	0.7898
(Zivkovic, 2004)	0.3033	0.8182	0.7383	0.7383	0.7938
(Kaewtrakulpong and Bowden, 2002)	0.0569	0.6480	0.4315	0.6743	0.7579
(Zivkovic and Heijden, 2006)	0.5904	0.8244	0.8828	0.7533	0.7068
proposed	0.8459	0.9100	0.9332	0.6719	0.8245

frames are used as a learning phase and are followed by 100 frames to which hand segmented groundtruth images are available for the evaluation. The dataset features typical underwater challenges like blur, haze, color attenuation, caustics and marine snow which all complicate the background modeling process.

A comparison between the proposed algorithm, the original GSM and other background subtraction algorithms is given in Table 1 and 2. It shows that the extended GSM is a substantial improvement to the original GSM on each of the five videos and also outperforms the other methods on the whole dataset. In Figure 6 some results of our algorithm for each of the five videos are depicted.

In the *Fish Swarm* video we could achieve the largest improvement, mainly because of the pre-segmentations which enabled us to build a far better background model of that scene. The main problem in this video is that there are always fishes in the middle of the scene which are also all quite similar to each other as well as the background. Therefore, a normal background modeling algorithm would take the fishes as part of the background and only the exclusion of moving objects from the updating process with the pre-segmentations could rectify that (see Figure 4).

Nonetheless, not all fishes in the *Fish Swarm* video could be detected since some of them barely move or are almost indistinguishable from the background. In the other four videos of the dataset the fishes can be detected very reliable by the proposed approach and the problems there mostly consist of false detections of shadows caused by the fishes or caustics on the water surface. It is a complicated task to avoid these errors since the algorithm needs to be very sensitive to detect fishes even when they are similar to the background which then causes these false detections.

5 CONCLUSION

In this paper we have enhanced the GSM background modeling by combining it with the Mixture of Gaussian idea and adding a foreground model. The foreground model is especially useful in scenes with swarms of fishes since the foreground objects in these scenes are all similar and can therefore be modeled accurate without a long adaption phase. Furthermore, we have used a coarse segmentation derived by the Flux Tensor to mark areas with possible foreground objects so that they can be excluded from the updating process of the background model. With this method we have generated more accurate background models without artefacts from foreground objects and hence could create better segmentations.

To include a spatial component we used the N^2Cut to adapt the segmentation to the smoothness of natural images and also corrected single false detection due to noise. We evaluated the proposed method on the Underwater Change Detection dataset to test it in these difficult situations and in scenarios with many foreground objects that are permanently visible. Especially on the crowded scenes the algorithm showed great improvements compared to other methods be-

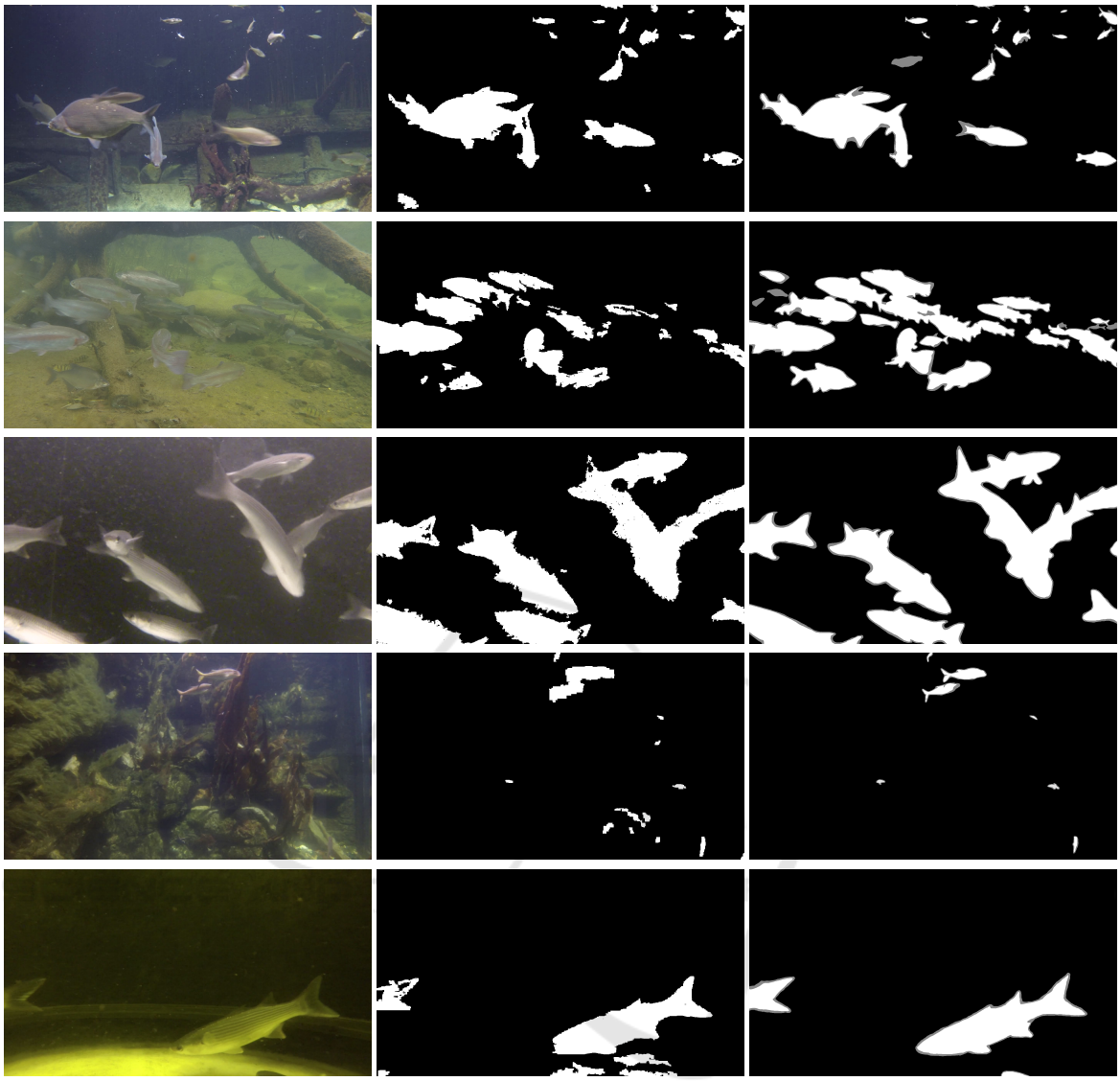

Figure 6: One frame of each of the five videos of the Underwater Change Detection dataset. From top to bottom are shown the videos: *Marine Snow*, *Fish Swarm*, *small Aquaculture*, *Caustics* and *two Fishes*. In the middle column is the segmentation of the proposed approach and in the right the ground truth data.

cause of the pre-segmentations but also on the other videos a continuous improvement to the normal GSM could be achieved.

REFERENCES

Barnich, O. and Droogenbroeck, M. V. (2011). Vibe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image Processing*, 20(6):1709–1724.

Benfold, B. and Reid, I. (2011). Stable multi-target tracking in real-time surveillance video. In *CVPR*, pages 3457–3464.

Bianco, S., Ciocca, G., and Schettini, R. (2015). How far can you get by combining change detection algorithms? *CoRR*, abs/1505.02921.

Bucak, S., Gunsel, B., and Guersoy, O. (2007). Incremental nonnegative matrix factorization for background modeling in surveillance video. In *Signal Processing and Communications Applications, 2007. SIU 2007. IEEE 15th*, pages 1–4.

Bunyak, F., Palaniappan, K., Nath, S. K., and Seetharaman, G. (2007). Flux tensor constrained geodesic active contours with sensor fusion for persistent object tracking. *J. Multimedia*, 2(4):20–33.

Gardos, T. and Monaco, J. (1999). Encoding video images using foreground/background segmentation. US Patent 5,915,044.

Hu, Z., Wang, Y., Tian, Y., and Huang, T. (2011). Selective eigenbackgrounds method for background subtraction in crowed scenes. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 3277–3280.

Kaewtrakulpong, P. and Bowden, R. (2002). An improved adaptive background mixture model for real-time tracking with shadow detection. In *Video-based surveillance systems*, pages 135–144. Springer.

Marghes, C., Bouwmans, T., and Vasiu, R. (2012). Background modeling and foreground detection via a reconstructive and discriminative subspace learning approach. In *Image Processing, Computer Vision, and Pattern Recognition (IPCV'12), The 2012 International Conference on*, volume 02, pages 106–112.

Mignotte, M. (2010). A label field fusion bayesian model and its penalized maximum rand estimator for image segmentation. *IEEE Transactions on Image Processing*, 19(6):1610–1624.

Nath, S. and Palaniappan, K. (2005). Adaptive robust structure tensors for orientation estimation and image segmentation. *Lecture Notes in Computer Science (ISVC)*, 3804:445–453.

Radolko, M., Farhadifard, F., Gutzeit, E., and von Lukas, U. F. (2015). Real time video segmentation optimization with a modified normalized cut. In *Image and Signal Processing and Analysis (ISPA), 2015 9th International Symposium on*, pages 31–36.

Radolko, M., Farhadifard, F., Gutzeit, E., and von Lukas, U. F. (2016). Dataset on underwater change detection. In *OCEANS 2016 - MONTEREY*, pages 1–8.

Radolko, M. and Gutzeit, E. (2015). Video segmentation via a gaussian switch background-model and higher order markov random fields. In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications Volume 1*, pages 537–544.

Ridder, C., Munkelt, O., and Kirchner, H. (1995). Adaptive background estimation and foreground detection using kalman-filtering. In *Proceedings of International Conference on recent Advances in Mechatronics*, pages 193–199.

Schindler, K. and Wang, H. (2006). Smooth foreground-background segmentation for video processing. In *Proceedings of the 7th Asian Conference on Computer Vision - Volume Part II*, ACCV'06, pages 581–590, Berlin, Heidelberg. Springer-Verlag.

Shelley, A. J. and Seed, N. L. (1993). Approaches to static background identification and removal. In *Image Processing for Transport Applications, IEE Colloquium on*, pages 6/1–6/4.

St-Charles, P. L., Bilodeau, G. A., and Bergevin, R. (2015). Subsense: A universal change detection method with local adaptive sensitivity. *IEEE Transactions on Image Processing*, 24(1):359–373.

Stauffer, C. and Grimson, W. (1999). Adaptive background mixture models for real-time tracking. In *Proceedings 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Vol. Two*, pages 246–252. IEEE Computer Society Press.

Toyama, K., Krumm, J., Brumitt, B., and Meyers, B. (1999). Wallflower: Principles and practice of background maintenance. In *Seventh International Conference on Computer Vision*, pages 255–261. IEEE Computer Society Press.

Wang, R., Bunyak, F., Seetharaman, G., and Palaniappan, K. (2014). Static and moving object detection using flux tensor with split gaussian models. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 420–424.

Warfield, S. K., Zou, K. H., and Wells, W. M. (2004). Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. *Ieee Transactions on Medical Imaging*, 23:903–921.

Wren, C., Azarbayejani, A., Darrell, T., and Pentland, A. (1997). Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:780–785.

Zivkovic, Z. (2004). Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 2 - Volume 02*, ICPR '04, pages 28–31, Washington, DC, USA. IEEE Computer Society.

Zivkovic, Z. and Heijden, F. (2006). Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recogn. Lett.*, 27(7):773–780.