On the segmentation and classification of water in videos

Pascal Mettes[1], Robby T. Tan[1], and Remco Veltkamp[1]
[1]*Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, Utrecht, the Netherlands*
P.S.M.Mettes@uva.nl, r.t.tan@uu.nl, R.C.Veltkamp@uu.nl

Keywords: hybrid water descriptor, mode subtraction, decision forests, markov random field, novel database

Abstract: The automatic recognition of water entails a wide range of applications, yet little attention has been paid to solve this specific problem. Current literature generally treats the problem as a part of more general recognition tasks, such as material recognition and dynamic texture recognition, without distinctively analyzing and characterizing the visual properties of water. The algorithm presented here introduces a hybrid descriptor based on the joint spatial and temporal local behaviour of water surfaces in videos. The temporal behaviour is quantified based on temporal brightness signals of local patches, while the spatial behaviour is characterized by Local Binary Pattern histograms. Based on the hybrid descriptor, the probability of a small region of being water is calculated using a Decision Forest. Furthermore, binary Markov Random Fields are used to segment the image frames. Experimental results on a new and publicly available water database and a subset of the DynTex database show the effectiveness of the method for discriminating water from other dynamic and static surfaces and objects.

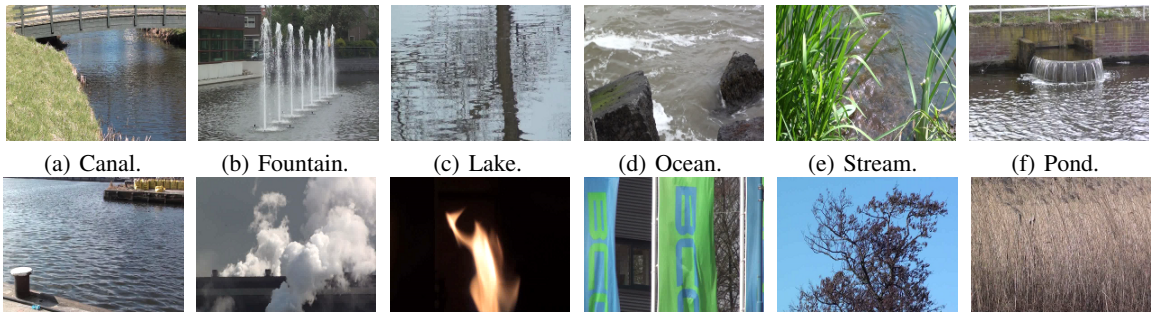
1 INTRODUCTION

Water recognition is a seemingly effortless task for humans, which is hardly surprising given the biological importance of water. While recent studies have indeed shown that humans are experts at such tasks (Sharan et al., 2013), there is little empirical knowledge on how water can be optimally recognized. Perhaps the most illustrative insight is provided in the work of Schwind, which indicates the importance of polarizing light reflected from water surfaces (Schwind, 1991). The experiments performed on water insects such as bugs and beetles showed that they are attracted by the horizontally polarized light from the reflections of water surfaces. However, the task of recognizing water from only images and videos, which do not possess polarization information, is still hardly problematic for human observers.

Therefore, the method provided here attempts to do the same, namely, to recognize water based only on visual appearance. The specific task of water identification in videos has, to the best of our knowledge, not been tackled on the scale presented in this work. Besides attempting to gain empirical knowledge, there is a wide range of applications that can benefit from such an algorithm, including: (inter-planetary) exploration, dynamic background removal, robotics, and aerial video analysis.

Traditionally, automatic water recognition is studied from two perspectives, namely as part of larger recognition tasks such as material recognition (Sharan et al., 2013; Hu et al., 2011; Varma and Zisserman, 2005) or dynamic texture recognition (Chan and Vasconcelos, 2008; Fazekas and Chetverikov, 2007; Saisan et al., 2001; Zhao and Pietikäinen, 2007), and in specialized and restricted environments, including autonomous driving systems (Rankin and Matthies, 2006) and maritime settings (Smith et al., 2003). Most current works in material and texture recognition are based on the hypothesis that target classes can be discriminated using distributions of local image features, global motion statistics, or learned ARMA models. Although interesting results have been reported, the descriptors themselves are generic and there is little analysis on which features work well for a certain texture. Furthermore, the approaches are usually global, which means that they are not directly applicable to localization tasks. On the other hand, water detection systems in autonomous driving systems and in maritime settings make explicit and non-generalizable assumptions, such as horizon location, camera height and orientation, and sky-water positioning, making the methods incapable of water detection from a broad scope.

Given the limitations of existing methods, a novel method is proposed in this paper. At the core of the



(a) Canal. (b) Fountain. (c) Lake. (d) Ocean. (e) Stream. (f) Pond.

(g) River. (h) Non-water.

Figure 1: Exemplary frames of categories in the water database.

method is a hybrid descriptor based on local spatial and temporal information. First, the input video is pre-processed to remove the background reflections and water colour, leaving only the residual image sequence, which conveys the motion of water. Given the residual images, local descriptors are extracted and classified using a Decision Forest (Bochkanov, 2013; Criminisi et al., 2012). The trained Forest can then be utilized to perform local classification for a collection of test sequences, where the probabilities are provided to a binary Markov Random Field to generate a segmentation for the frames in the test sequences based on the presence or absence of water.

Since this work is specifically aimed at water detection and given the supervised nature of the algorithm, a second contribution of this work is the introduction of a new database. The water database, further elaborated at the end of this section, consists of a set of complex natural scenes with a wide variety of water surfaces. Experimental evaluation on this database and on the DynTex database (Péteri et al., 2010) shows the effectiveness of the proposed method for both video classification and spatio-temporal segmentation.

The paper is organized as follows. This Section is concluded with an elaboration of the water database, while Section 2 discusses the works related to computational water recognition. Section 3 provides an analysis of the temporal and spatial behaviour of water, the descriptors derived from that analysis, and the process of local classification. This is followed by the global segmentation step in Section 4. Section 5 shows the experimental evaluation on the databases and the paper is concluded in Section 6.

1.1 Water Database

As stated above, experimental evaluation is performed in this work on a novel database, in order to tackle the challenge problem of water detection[1]. Although databases used in dynamic texture recognition do contain water videos (Péteri et al., 2010), they do not contain water videos in the quantity and variety desired for water recognition and localization. The novel water database contains a set of positive and negative videos (i.e. water and non-water videos), from a wide variety of natural scenes. In total, the database consists of 260 videos, where each video contains between 750 and 1500 frames, all with a frame size of 800×600. The positive class consists of 160 videos of predominantly 7 scenes; oceans, fountains, ponds, rivers, streams, canals, and lakes. The negative class on the other hand can be represented by any other scene. In the database, categories with seemingly similar spatial and temporal characteristics are chosen, including trees, fire, flags, clouds/steam, and vegetation. Examples of the categories of the database are shown in Fig. 1. Since the focus of this work is aimed at characterizing the behaviour of water, camera motion is considered a separate problem and it is therefore not integrated into the database.

2 RELATED WORK

Original work on material classification attempted to discriminate materials photographed in laboratory settings, e.g. based on distributions of Gabor filter responses (Varma and Zisserman, 2005) or image patch exemplars (Varma and Zisserman, 2009). More recently, the classification problem has shifted to the real-world domain with the Flickr Materials Database (Sharan et al., 2009), which includes water as one of the target classes. Current approaches attempt to discriminate materials based on concatenations of local image feature distributions (Hu et al., 2011; Sharan et al., 2013). Although it is possible to apply these ap-

[1]The database is available for download at http://www.staff.science.uu.nl/˜tan00109/

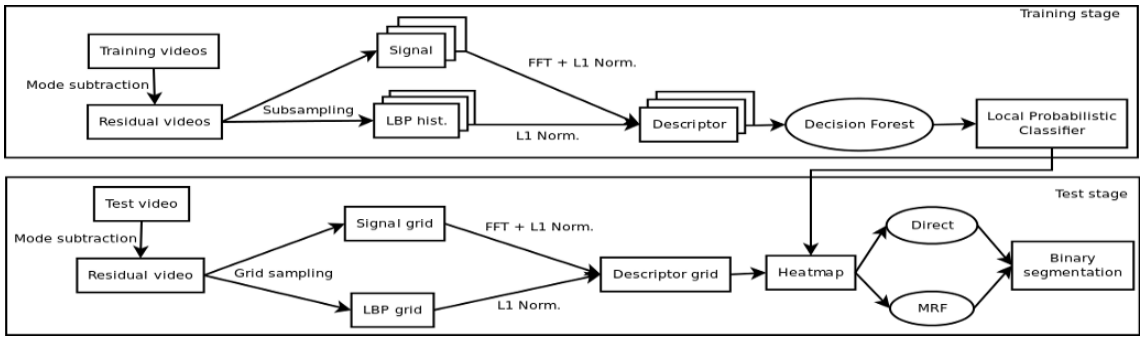


Figure 2: An overview of the water segmentation method, where the test stage is done for all frames of the test videos.

proaches to the more specific problem of water recognition, they are severely restricted in multiple aspects. First, the algorithms are concerned with classification solely from images, so any form of temporal information is neglected. Second, and more importantly, the algorithms provide a black-box process, where it is unknown how or why certain materials (such as water) can be discriminated using the recommended image features (e.g. SIFT, Jet, Colour, etc.).

A field more directly related to the problem tackled in this work is dynamic texture recognition, which is roughly dominated by two approaches. The first approach is the discrimination of dynamic textures based on two-frame motion estimation. Generally, well-known conventional optical flow methods are used for dense motion estimation, after which classification is performed based on global motion statistics. Global statistics include the curl and divergences of the flow field, and the probability of having a characteristic flow direction and magnitude (Fazekas and Chetverikov, 2007; Péteri and Chetverikov, 2005). Although high recognition rates have been reported for such methods, the use of conventional optical flow is problematic for water in natural scenes. In general, water does not meet the conditions of optical flow (Beauchemin and Barron, 1995). Another pressing problem is that flow-based methods are focused on classification, not segmentation.

A second dominant approach is the global modeling of videos using Linear Dynamical Systems (LDS) (Chan and Vasconcelos, 2008; Doretto et al., 2003; Mumtaz et al., 2013; Saisan et al., 2001). In its essence, LDS is a latent variable model which projects video frames to a lower dimensional space and tracks the temporal behaviour in that lower dimensional space. The use of LDS in dynamic texture recognition was first popularized by the work of (Saisan et al., 2001), mostly due to the proposed relatively efficient parameter learning procedure and the encouraging classification results. Although the use of LDS is intuitively appealing, the original formulation of LDS is limited, since it cannot handle multiple objects/textures in a single video. Furthermore, little investigation has been done as to which textural elements can be captured with LDS. Multiple extensions have been made to handle the presence of multiple textures, but in current literature, LDS is used either as a segmentation or classification method, but not as a joint segmentation-classification problem (i.e. dividing the pixels into different coherent parts and classifying each part, as is done in this work).

3 LOCAL WATER DETECTION

The primary focus of this work is the generation of local descriptors based on the analysis of the spatio-temporal behaviour of water. Here, both a temporal and spatial descriptor are presented which are distinctive enough for direct identification of water surfaces on a local scale. A generalized overview of the algorithm is shown in Fig. 2. First, the videos are pre-processed to increase invariance to water colours and reflections. After that the temporal and spatial descriptors are extracted and used as feature vectors for a Decision Forest. Given a test video, the local descriptors are extracted and classified using the trained Forest. The probability outputs are then used as input for a regularization step using spatio-temporal Markov Random Fields. In this section, the analysis and descriptor generation is provided, as well as the probabilistic classification.

3.1 Creating residuals

A major aspect of water surfaces is the inherent variability they possess, due to water colour, ripples, background reflections, weather conditions, etc. Rather than trying to exploit dominant features due to water colour or background reflections, the focus of this work is to generate features which are invariant to these aspects, and in effect state something about

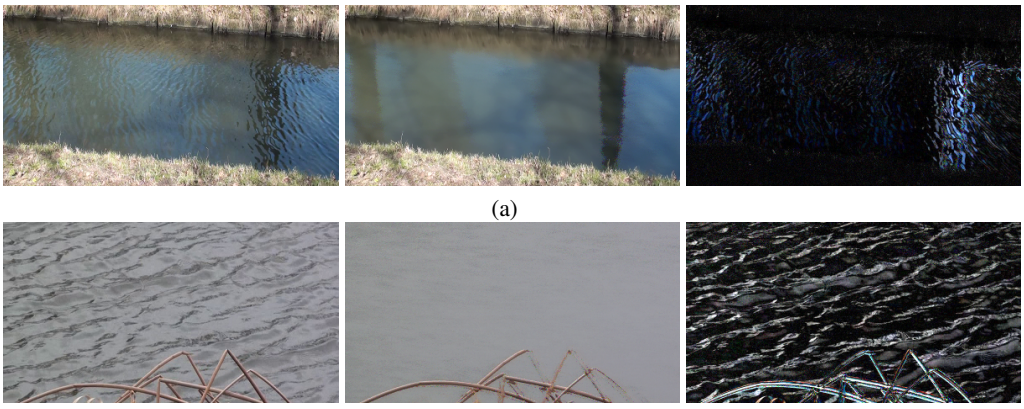


Figure 3: A typical frame, temporal mode, and residual for 2 videos.

the general nature of water surfaces. This is done by first obtaining the water colour as well as background reflections, and then removing them, leaving only residual images.

The aim of the residual frames is to highlight water ripples, instead of reflections. This is realized by performing temporal mode subtraction for each pixel:

$$R_t(x,y)[c] = |I_t(x,y)[c] - M(x,y)[c]|, \qquad (1)$$

for each $c \in \{R,G,B\}$ separately, where $M(x,y)$ denotes the temporal mode of pixel (x,y). The temporal mode of a single pixel for a single colour channel can be computed as follows:

$$M(x,y)[c] = \max_p \sum_{t=1}^{T} \mathbb{1}\{I_t(x,y)[c] = p\}, \qquad (2)$$

where T denotes the total number of frames and $\mathbb{1}\{\cdot\}$ is the indicator function. A typical frame, temporal mode, and corresponding residual frame of 2 videos are shown in Fig. 3. Note that this simple procedure does not remove all reflections, but it can successfully find the most dominant elements of reflection, such that the residual frames highlight water ripples.

3.2 Local temporal descriptor

The idea behind the temporal water descriptor is that the water ripples indicate the motion characteristics. The motion characteristics of water are hypothesized to constitute several aspects. Most notably, it is hypothesized here that this type of motion is gradual and repetitive, since ripples re-occur at the same location over time. Other dynamic and static processes might partially share the local temporal properties of water, but not statistically to the same extent.

Based on the defined hypotheses, the next step is to create a descriptor which judges a local spatio-temporal patch on these hypotheses. To achieve this,

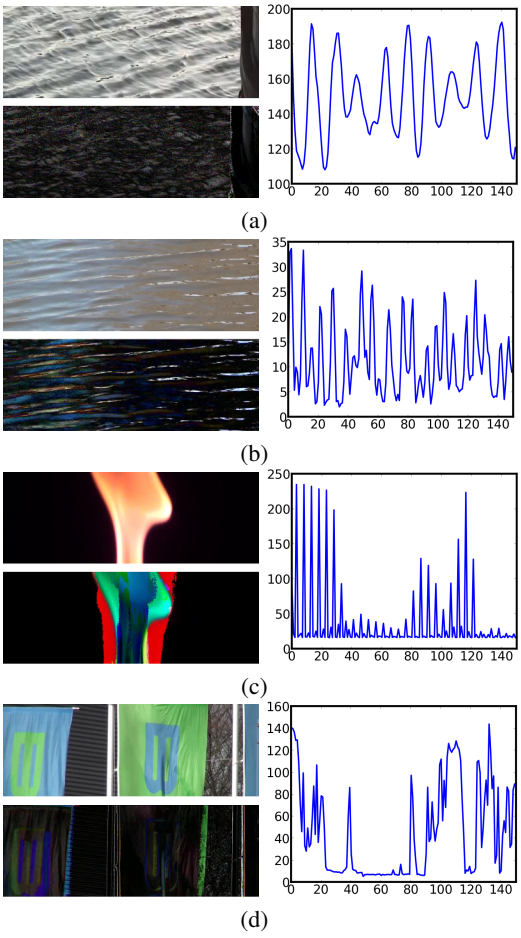
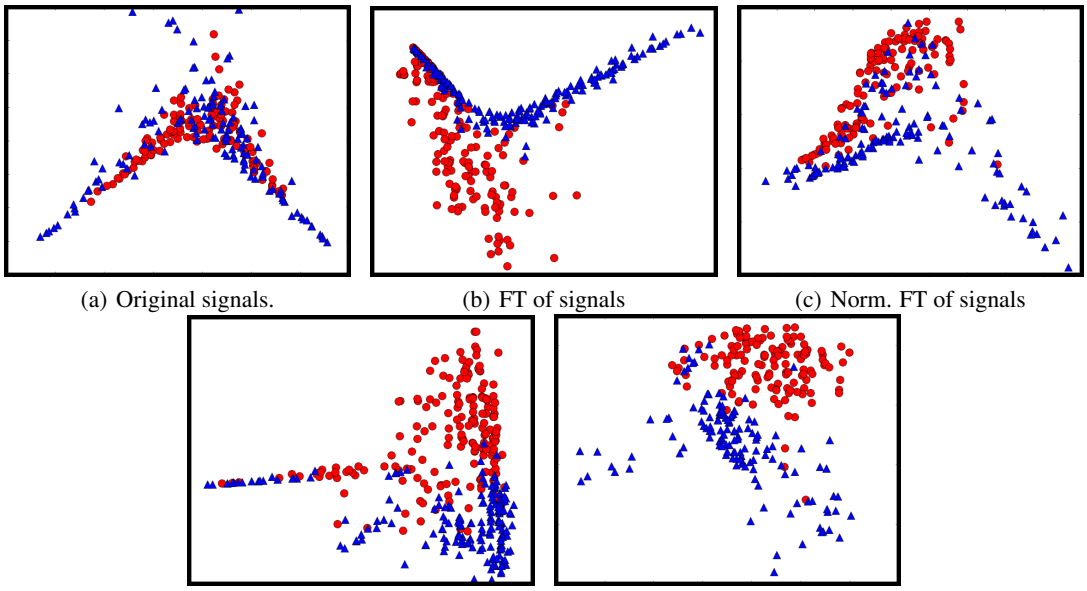


Figure 4: Four exemplary frames and sample signals.

an m-dimensional signal is first introduced, which represents the mean brightness value of an $n \times n$ patch for m consecutive frames at exactly the same location. The result is a list of brightness values which can be seen as a 1-dimensional signal. Fig. 4 shows mul-



(a) Original signals. (b) FT of signals (c) Norm. FT of signals

(d) LBP (e) Hybrid

Figure 5: Isomap projections of sample locations of trees (blue) and water (red). The Figure shows how using normalized Fourier Transforms improves separation using temporal information (a,b,c). More interesting, it clearly shows that fusing temporal and spatial information creates a further boost in separation (c,d,e).

tiple examples of water and non-water scenes, along with a mean brightness signal of a selected location. From the Figure, the initial thoughts regarding regularity and repetition are already clearly visible.

Given the m-dimensional signals, an immediate thought is to use them directly as the descriptor. The direct use of the signal itself as the descriptor is however erroneous, due to lack of invariance. Signals with similar levels of smoothness and regularity can have a large distance when comparing the signals directly. A descriptor based on the m-dimensional signal should in effect be invariant to temporal shifts, brightness shifts, and brightness amplitudes. Rather than creating a distance measure which explicitly enforces these types of invariance by means of normalization and distance recalculation for all possible temporal and brightness shifts, a descriptor is created here by extracting the signal characteristics using the 1-dimensional Fourier Transform. For an m-dimensional signal S, the corresponding Fourier Transform F is also m-dimensional, where each component $i \in \{1,m\}$ is computed as:

$$F_i = |\sum_{j=1}^{m} S_j e^{-2\pi i j\sqrt{-1}/m}|. \quad (3)$$

In other words, from the signal, an m-dimensional descriptor $[F_1,..,F_m]$ is computed. Although this descriptor is invariant to the temporal and brightness shifts, it is not invariant to brightness amplitudes. Therefore, the final temporal descriptor is generated by performing $L1$-normalization on the components, such that two signals are compared by the distribution of the Fourier Transform, i.e.:

$$F_i = \frac{|\sum_{j=1}^{m} S_j e^{-2\pi i j\sqrt{-1}/m}|}{\sum_{k=1}^{m}|\sum_{l=1}^{m} S_l e^{-2\pi k l\sqrt{-1}/m}|}. \quad (4)$$

A practical justification of using the $L1$-normalized Fourier Transform as the temporal descriptor is shown in Fig. 5. The Figure displays the Isomap projection (Tenenbaum et al., 2000) of samples taken from water and tree videos. An ideal local descriptor has linear separability in the original feature space. For visualization purposes, the projected feature space is used here, but the full feature space is used in the classification. As can be seen in Fig. 5(a) and Fig. 5(b), the original signals and their Fourier Transform are rather impractical in terms of classification, while a separation is clearly visible in Fig. 5(c), although there is an area of overlap.

The choice of signal length m is a trade-off. Ideally for recognition, m is equal to the total number of frames in the video, since the longer the signal, the less likely it is that non-water surfaces mimic the temporal behaviour of water. On the other hand, this approach makes it impossible to detect temporal discontinuities. Since the focus lies primarily on discriminations while still being able to detect obvious outliers, m is set to 200 frames here.

3.3 Local spatial descriptor

The above defined descriptor captures the temporal behaviour of water, but ignores the local spatial information, most notably the spatial layout of water waves and ripples. Given that water waves, ripples, and fountains are highly deformable, a descriptor is desired which provides spatial information on a local patch without requiring an explicit model of water waves. To meet this desire, the local spatial characteristics of water surfaces are extracted using Local Binary Pattern histograms (Zhao and Pietikäinen, 2007). For a single pixel, the Local Binary Pattern is computed by comparing the grayscale values of the pixel to a set of local spatial neighbours. In this work, the 8 direct neighbours of a pixel are used for comparison. As such, the Local Binary Pattern value of a single pixel is computed as:

$$LBP_{8,1}(g^c) = \sum_{p=0}^{7} H(g_p^c - g^c)2^p, \tag{5}$$

where g^c denotes the center pixel for which the LBP value is computed, $\{g_i^c\}_{i=0}^7$ denotes the set of direct neighbours of g^c, and $H(\cdot)$ is the well-known discrete Heaviside step function, defined as:

$$H(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0. \end{cases} \tag{6}$$

Since 8 neighbours are used in the comparison, the corresponding LBP value can take $2^8 = 256$ values. In order to compute a LBP histogram of a local patch, the LBP values of the pixels in the patch are computed and placed in their corresponding integer bins of the 256-dimensional histogram. The resulting histogram is normalized afterwards.

As stated above, a primary justification for the use of Local Binary Pattern histograms as a spatial descriptor is due to the pseudo-orderless nature of the descriptor, which means that water waves do not need to be modeled explicitly. Furthermore, the high dimensionality of the histograms provide desirable discrimination abilities. Similar to the temporal descriptor, the practical validity of the LBP histograms can be shown by examining the projected feature space. The early fusion (Snoek et al., 2005) of the temporal and spatial descriptors into a hybrid descriptor, results in a feature space where its projection is almost nearly linearly separable for the randomly selected local patch, as can be seen in Fig. 5(e). The importance of a pseudo-orderless spatial descriptor came to light after the investigations into explicit water modeling turned out to be impractical. This conclusion is consistent with literature on dynamic texture recognition. For example in overlapping work of Zhao and Pietkäinen, multiple extensions of LBP have been proposed, such as VLBP (Zhao and Pietikäinen, 2006) and LBP-TOP (Zhao and Pietikäinen, 2007). VLBP is however impractical for local direct identification, since each histogram would have a length of 2^{14} or even 2^{26}, due to the fact that both spatial and temporal neighbours need to be compared against the central pixel. Similar statements can be made regarding LBP-TOP. Therefore, the original purely spatial LBP descriptor is used here.

3.4 Probabilistic classification

Now that the behaviour of water on a local temporal and spatial level have been defined, the next step is to exploit the descriptors for probabilistic classification. Contrary to computing distributions of descriptors as is usual in global classification tasks, the descriptors are used directly as feature vectors for probabilistic classification using Decision Forests. In the training stage of the algorithm, the descriptors are extracted from the training videos and used as feature vectors for the Decision Forest. Since the number of patches per frame can be considerably large, selecting patches from a uniform grid for each frame of each training video is undesirable, given the amount of time required for training. For that reason, a random sampling approach is employed by selecting a small number of patches per frame per training video.

Given the use of random sampling, roughly 2500 local patches are selected per training video. The 456-dimensional feature vectors for the patches of all training videos are then fed to the Decision Forest for probabilistic classification. The primary parameters of the forest - the number of trees and the randomness factor - can be set using cross-validation.

In the testing stage, descriptors need to be extracted from all parts of the frames of the test videos. Therefore patches are extracted from a dense rectangular grid. The descriptors yielded from the grid are individually given to the trained forest, yielding a matrix of probability outputs, which can be seen as a heatmap. A major advantage of direct local classification is that each local part of the video is classified separately. However, since this approach yields a great number of separate classifications, it can be expected that multiple miss-classifications occur within and between the frames of a test video. For that reason, a last step of this algorithm is the use of Markov Random Fields for spatio-temporal regularization.

4 HEATMAP REGULARIZATION

The additional information on the probabilistic (un)certainty of classified local patches, instead of direct decisions, opens up the possibility to discrete optimization on the heatmaps. The discrete optimization takes the form in this work of a Markov Random Field (Boykov and Kolmogorov, 2004), which serves as a regularization step. More formally, the optimization problem of the MRF can be stated as a minimization problem with the following objective function:

$$f(x) = \sum_{p \in V} V_p(x_p) + \lambda \sum_{(p,q) \in C} V_{pq}(x_p, x_q), \quad (7)$$

with V the elements of the heatmap, and C the set of all cliques. The first term of the objective function - the data term - is then defined as:

$$V_p(x_p) = \begin{cases} 1 - M_p & \text{if } x_p \text{ is water} \\ M_p & \text{otherwise} \end{cases} \quad (8)$$

where M_p denotes the probability of node (i.e. heatmap pixel) p of begin water. The second term - the prior term - is defined such that different labels within cliques are penalized:

$$V_{pq}(x_p, x_q) = |x_p - x_q|, \quad (9)$$

given that the label water is defined as 1 and the label non-water is defined as 0.

An interesting element within the MRF formulation are the cliques. Rather than only enforcing similarity between neighbouring pixels in a single heatmap, a form of temporal regularization is also desired, since water location is not expected to change sharply over time. Therefore, a spatio-temporal Markov Random Field formulation is opted here, where each element of the heatmap is connected to both its 4 spatial neighbours and 2 temporal neighbours.

An important element in the minimization procedure is the relative weight of the probabilities (data term) and spatial consistency (prior term), denoted by λ in Eq. 7. For a low value for λ, the individual probabilities are deemed important, resulting in a segmentation with a lot of detail, but also with outliers. On the other hand, a high value for λ results in a segmentation with little outliers, at the cost of loss of detail at borders between water and non-water regions. The influence of the λ term is evaluated in Section 5. Since not all pixels on a single frame are classified, the binarized heatmap only contains the segmentation result for a subset of the pixels on a rectangular grid. The segmentation results are therefore bi-cubicly interpolated such that each pixel is classified as either being water or non-water.

5 EXPERIMENTAL EVALUATION

The effectiveness of the algorithm presented in the previous sections is validated on the novel water database by examining both the segmentation quality (i.e. the classification of each pixel of each frame) and the classification quality (i.e. the classification of a whole video with a binary mask). In the implementation of the algorithm, the videos of the database are randomly split with a 60/40 ratio into a train and test set. For each test video, the segmentation is computed for 250 frames.

In the evaluation, the segmentation fit of the video is defined as the average of the fit of the individual segmentations with the supplemented binary mask. Formally, the segmentation fit S of a segmented video V compared to a mask M is computed as:

$$S(V, M) = \frac{\sum_{i=1}^{|V|} s(V_i, m)}{|V|} \times 100\%, \quad (10)$$

where $|V|$ denotes the number of frames in V, V_i denotes the i^{th} frame, and $s(V_i, M)$ is defined as:

$$s(V_i, M) = 1 - \frac{\sum_{x=1}^{W} \sum_{y=1}^{H} |V_i[x,y] - m[x,y]|}{W \times H}, \quad (11)$$

with W and H for resp. the width and height of the video and the pixel values of the segmentations and the mask are 1 for water and 0 for non-water. Given a set of segmentations and a binary mask, the whole video can be classified as water if the ratio of water pixels in the mask region is at least a half, and non-water otherwise.

In Table 1, a numerical overview is provided of the averaged segmentation fit for the individual and combined descriptors, where the algorithm is performed on multiple random splits. From the Table, it is clear that both the temporal and spatial descriptors are able to robustly segment video frames based on the presence or absence of water. Furthermore, the combination of the descriptors into a hybrid descriptor has a strictly positive influence on the segmentation quality. A similar statement can be made regarding the regularization step, where the combination of the hybrid descriptor and the spatio-temporal Markov Random Field yield an average segmentation fit of 93.19%. In Figure 6, exemplary binary segmentations are shown for complex test videos in the database.

Descriptor	No MRF	ST-MRF, $\lambda = 1.0$
Hybrid	$90.38\% \pm 0.5\%$	$\mathbf{93.19\% \pm 0.2\%}$
LBP	$85.95\% \pm 2.4\%$	$90.16\% \pm 2.3\%$
Temporal	$83.42\% \pm 1.3\%$	$87.42\% \pm 0.8\%$

Table 1: Segmentation quality of the descriptors.

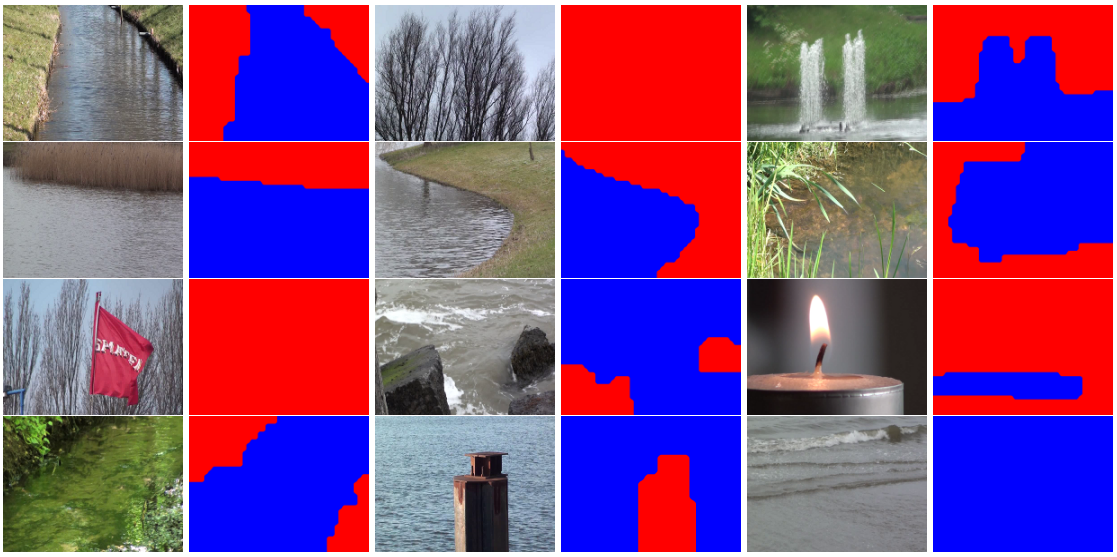


Figure 6: Exemplary segmentations yielded for the water database.

An influential parameter is the value for λ in the regularization step. In Fig. 7, the segmentation fit as a function of the parameter is shown. From the Figure, the trade-off between dependence on the classification result and spatial coherence in the MRF is evident. The result of Table 1 are yielded with 200 frames per descriptor and $\lambda = 1.0$, since the segmentations at that value on average prove to be optimal with respect to both the individual classification and spatial coherence. In Fig. 8, the effect of regularization is visualized for a water and a non-water video. With regularization, the ST-MRF makes sure that a boundary between water and non-water regions is only created if there is enough support over the whole image plane.

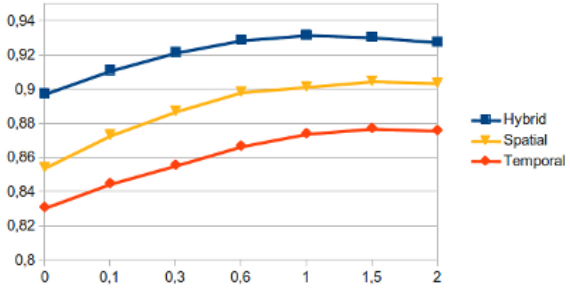
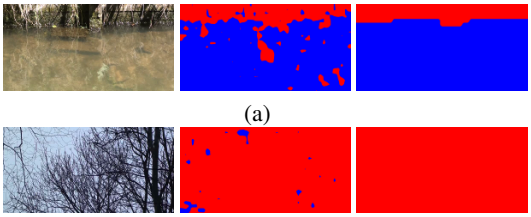


Figure 7: Segmentation fit as a function of λ.

An overview of the results of binary video classification is shown in Table 2. The video classification problem - less informative than the segmentation problem - is not only performed for the algorithm of this work, but also for multiple algorithms from related fields. These algorithms include the spatio-temporal extensions of LBP (Zhao and Pietikäinen, 2007), LDS (Saisan et al., 2001), Gabor filter responses (Varma and Zisserman, 2005), and optical flow statistics (Péteri and Chetverikov, 2005).



(a)

(b)

Figure 8: Two examples of the merits of using regularization (right column) over direct classification (middle column).

Classification method	Recognition rate
Our method, hybrid descriptor	**96.5% ± 0.6%**
Our method, LBP histogram	94.2% ± 1.4%
Our method, temporal descriptor	91.9% ± 0.8%
Volume LBP	93.5% ± 1.2%
LBP-TOP	93.3% ± 0.8%
MR8 Filter Bank distributions	80.8% ± 6.7%
Linear Dynamical Systems	71.2% ± 2.8%
Horn-Schunck, 4 flow statistics	65.7% ± 2.0%
Lucas-Kanade, 4 flow statistics	61.2% ± 3.4%

Table 2: Overview of the recognition rates for binary video classification.

The results from Table 1 and 2 indicate that the spatial and temporal descriptor can effectively capture the local behaviour of water. The hybrid descriptor outperforms well known generic algorithms from dynamic/static texture recognition for the more specific problem of water classification. A final best result

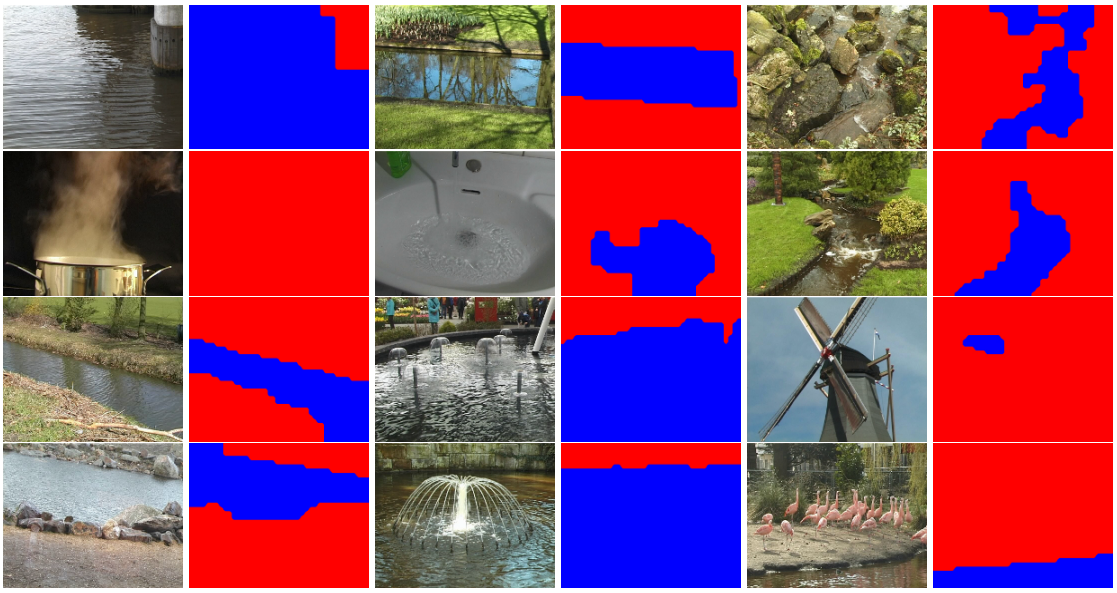


Figure 9: Exemplary segmentations yielded for the DynTex database.

is achieved with the hybrid descriptor and ST-MRF, with a segmentation fit of **93.19%** and a video classification rate of **96.47%**. A primary reason for the overall high recognition rates in this work is because this is a binary problem, i.e. if a patch of a tree is classified as fire, it is correct, since the water/non-water line is not crossed. Also, the classification results are generally higher, since a perfect segmentation result is not required to yield a correct overall classification.

Method	Segmentation	Classification
Ours, hybrid	**92.7%**	**100%**
Ours, temporal	85.0%	87.5%
Ours, spatial	81.3%	87.5%
VLBP	-	90.0%
LBP-TOP	-	87.5%
LDS	-	75.0%
MR8	-	72.5%
HS flow	-	57.5%
LK flow	-	55.0%

Table 3: Results on the DynTex subset.

In order to emphasize the effectiveness of the descriptors, the algorithm is also run on videos in the DynTex database (Péteri et al., 2010). Since only a part of the DynTex database contains water surfaces, a subset of 80 water and non-water videos have been selected for evaluation. A second motive for experimenting on the DynTex database is that it provides a comparison for water detection against other non-water textures and objects, such as humans, animals, traffic, windmills, flowers, and cloths. For the classification process, the 80 videos are split into a trainset and testset of 40 videos, while the trainset is complemented with an additional 100 videos from the water database. Exemplary segmentations are shown in Fig. 9. The numerical results for the segmentations and classifications are provided in Table 3. The results of Table 3 further indicate the effectiveness of the algorithm.

6 CONCLUSIONS

In this work, a method and database are introduced for the spatio-temporal identification of water surfaces in videos. Rather than tackling water detection as an instance of a more generic detection method, such as materials recognition or dynamic texture recognition, this method attempts to recognize water based on the specific spatial and temporal behaviour of water surfaces. Experimental evaluation on a novel water detection database shows the efficiency of the method, outperforming well-known existing algorithms from static and dynamic texture recognition.

In future work, the algorithms presented here can be used to tackle the problem of real-time water detection with moving cameras. Real-time detection can be investigated by creating a parallel implementation of the feature extraction and classification of the local descriptors. Given that regularization is currently a post-processing step, it should be incorporated in the classification stage in a real-time setting, e.g. using Kontschieder et al.'s recently introduced Geodesic Forests (Kontschieder et al., 2013).

ACKNOWLEDGEMENTS

This research is supported by the FES project COMMIT. Furthermore, we would like to thank Renaud Péteri for providing access to the DynTex database.

REFERENCES

Beauchemin, S. and Barron, J. (1995). The computation of optical flow. *ACM Computing Surveys*, 27(3):433–466.

Bochkanov, S. (1999-2013). Alglib software library (www.alglib.net).

Boykov, Y. and Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*.

Chan, A. and Vasconcelos, N. (2008). Modeling, clustering, and segmenting video with mixtures of dynamic textures. *PAMI*, 30(5):909–926.

Criminisi, A., Shotton, J., and Konukoglu, E. (2012). Decision forests. *Foundations and Trends in Computer Graphics and Vision*, 7(2):81–227.

Doretto, G., Cremers, D., Favaro, P., and Soatto, S. (2003). Dynamic texture segmentation. *ICCV*, 2:1236–1242.

Fazekas, S. and Chetverikov, D. (2007). Analysis and performance evaluation of optical flow features for dynamic texture recognition. *SPIC*, 22:680–691.

Hu, D., Bo, L., and Ren, X. (2011). Toward robust material recognition for everyday objects. *BMVC*, pages 48.1–48.11.

Kontschieder, P., Kohli, P., Shotton, J., and Criminisi, A. (2013). Geof: Geodesic forests for learning coupled predictors. *CVPR*.

Mumtaz, A., Coviello, E., Lanckriet, G., and Chan, A. (2013). Clustering dynamic textures with the hierarchical em algorithm for modeling video. *PAMI*, 35(7):1606–1621.

Péteri, R. and Chetverikov, D. (2005). Dynamic texture recognition using normal flow and texture regularity. *PRIA*, 3523:223–230.

Péteri, R., Fazekas, S., and Huiskes, M. (2010). Dyntex: A comprehensive database of dynamic textures. *Pattern Recognition Letters*, 31(12):1627–1632.

Rankin, A. and Matthies, L. (2006). Daytime water detection and localization for unmanned ground vehicle autonomous navigation. *Proceeding of the 25th Army Science Conference*.

Saisan, P., Doretto, G., Wu, Y. N., and Soatto, S. (2001). Dynamic texture recognition. *CVPR*, 2:II–58–II–63.

Schwind, R. (1991). Polarization vision in water insects and insects living on a moist substrate. *Journal of Comparative Physiology A*, 169(5):531–540.

Sharan, L., Liu, C., Rosenholtz, R., and Adelson, E. (2013). Recognizing materials using perceptually inspired features. *IJCV*, pages 1–24.

Sharan, L., Rosenholtz, R., and Adelson, E. (2009). Material perception: What can you see in a brief glance? [abstract]. *Journal of Vision*, 9(8):784.

Smith, A., Teal, M., and Voles, P. (2003). The statistical characterization of the sea for the segmentation of maritime images. *Video/Image Processing and Multimedia Communications*, 2:489–494.

Snoek, C., Worring, M., and Smeulders, A. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM.

Tenenbaum, J., de Silva, V., and Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Varma, M. and Zisserman, A. (2005). A statistical approach to texture classification from single images. *IJCV*, 62(1):61–81.

Varma, M. and Zisserman, A. (2009). A statistical approach to material classification using image patch exemplars. *PAMI*, 31(11):2032–2047.

Zhao, G. and Pietikäinen, M. (2006). Local binary pattern descriptors for dynamic texture recognition. *ICPR*, 2:211–214.

Zhao, G. and Pietikäinen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *PAMI*, 29(6):915–928.