Signal, Image and Video Processing (2019) 13:341–348
https://doi.org/10.1007/s11760-018-1362-1

ORIGINAL PAPER

Systematic approach to binary classification of images in video streams using shifting time windows

Adam Blokus[1] · **Henryk Krawczyk**[1]

Received: 7 September 2017 / Revised: 28 August 2018 / Accepted: 5 September 2018 / Published online: 17 September 2018

Abstract
Multiple algorithms classifying frames in video sequences consider them only as separate images. After pointing out the properties of real-life recordings and classifications of their frames, we propose a new shifting time window approach for improving binary classifications. It proceeds in two steps: First, well-known classification algorithms are used separately for each frame to acquire preliminary classifications. Secondly, the results of the previous step are analyzed in relatively short sequences of consecutive images (the shifting time window). Taking into account the continuous nature of analyzed real-life videos, the preliminary binary classification sequences can be corrected. In consequence, the classification quality is improved. Furthermore, we offer a systematic approach where all parameters of the proposed algorithm (such as the window length or vote weight distribution in the window) are considered and their optimal values are determined. Experiments on representative examples confirm the advantages of the proposed approach.

Keywords Image classification in videos · Binary classifications · Temporal relations · Continuous properties · OFA and FSA methods

1 Introduction

Having identified a number of algorithms which classify single images of a video stream, we can ask whether this approach is the optimal one in terms of real-life video sequences. While detecting a static feature requires only a single picture (frame), in real-life videos we might also consider a longer sequence of frames where the same feature should be visible.

Algorithms such as those presented in [1,2] tend to consider image classification in videos separately for each frame. Further on, we will denote this kind of algorithms as One Frame Analysis (OFA). Their main advantage is a relatively low cost of preparing image datasets and availability of well-established methods. The features they detect are static. For example, lesions in endoscopic videos or people in surveillance recordings can be recognized in single pictures, but their visibility can be expected to last for a number of consecutive frames.

✉ Adam Blokus
ablokus@eti.pg.gda.pl

[1] Faculty of Electronics, Telecommunication and Informatics, Gdańsk University of Technology, Gdańsk, Poland

The continuity of videos is implicitly the basis of object tracking methods which adaptively adjust the representations of tracked objects [3,4] or exploit their inter-frame similarity [5]. Even though an observed item can change dramatically over time, all changes are assumed to be gradual and traceable. Other methods, which track either faces [6] or hand movements [7], initially detect a trajectory in particular frames and smoothen it in a second step.

Various works can be singled out, which tackle particular aspects of considering the relationships between classifications of consecutive video frames. Such an approach can be applied in the frame-classifying method itself or by introducing an additional post-processing step to account for the temporal structure. The latter option will be further investigated within this paper.

For example, the authors of [8] proposed composing a scene segmentation algorithm of simpler classifying algorithms. The results are rationalized in terms of their temporal structure with a hidden Markov model.

Such an approach has also benefited the authors of [9], who have improved the recognition rate of text objects when temporal context in the video has been considered. Also [10] increased the recognition quality after introducing a shifting window post-processing step. In [11] outliers are corrected

in a final step, changing minor errors throughout sequences of 100 frames. Using small time windows to analyze ranked lists of identifications successfully improved the identification of pedestrians in [12]. The works introduced within this paragraph are the only ones which explicitly evaluate the influence of improving classifications by utilizing temporal properties of the video—which is the essence of our method. This information is presented in Sect. 5 for a comparison with the proposed approach.

Sliding windows have been used for video summarization in [13]. To establish relevant neighborhoods for key frames, the window size was adapted in respect to cuts of the video. The authors of [14] incorporate a post-processing step with a shifting window of size 5, where a majority vote determines the final classification of a frame (the segment of the gastrointestinal tract it represents). It's only briefly implied that this step improves the accuracy of their method.

In series of classifications, single outliers can be considered to be potential mistakes. A broad overview of outlier detection methods has been provided by Gupta et al. [15,16]. They compare a single value with its prediction based on its neighborhood (one-sided or two-sided). The prediction can be the median [17], mean [18] or a more complex function [18] of the values in the neighborhood. These methods correspond directly to the shifting window approach presented in [14,19,20], with the neighborhood equivalent to the shifting windows.

Summing up, multiple works can be pointed out where the neighborhood of classified frames has been taken into consideration. They have proven to be improving classification results, but so far there was no effort to propose and analyze a general approach.

The contributions introduced by this work are twofold. First, we have identified the need and motivation for introducing a method which allows to consider the temporal information, which is neglected by OFA methods. Those motivations are identified in the implicit use of temporal relations in multiple papers. They are also confirmed by a theoretical analysis. Secondly, the FSA method is proposed, which allows to improve existing OFA algorithms. It includes temporal information in a post-processing step, which allows to utilize all benefits of OFA methods. The experimental evaluation confirms the efficiency of the method and allows to understand how its controlling parameters affect it.

This paper considers primarily binary classifications, but a generalization for multi-categorical classifications can be conceived, e.g., for facial expressions [21]. This has been noted in the conclusions of this paper.

In the next section, we will continue with a brief discussion of how the continuity of real-life videos allows us to make assumptions regarding the structure of their classification sequences. As a conclusion, in Sect. 3 we propose a new scheme for improving algorithms which assign binary properties to video frames. The new method is evaluated according to a scheme presented in Sect. 4 on two real-life datasets as well as an artificial stream. The results of the evaluation are presented in Sect. 5. Finally, the conclusions of this paper are summarized in Sect. 6.

2 Continuity and change

The main context in which we consider the classified videos is their pace of change and what *continuous* features they therefore present (as discrete images of continuous processes).

We define the process of classification as a sequence of transformations between different spaces. The whole concept is presented in Fig. 1, where each subfigure corresponds to a step in transforming the recorded view into a sequence of consecutive classifications:

(a) The observed reality is of continuous nature, both in terms of the time axis and the changes it undergoes.
(b) The video recording represents the reality as closely as possible. Still, frames are recorded only in regular intervals (*step* value) and the acquired images are not fully accurate representations (observation axis—rounded values with a difference of Q).
(c) When every frame is classified separately, the classifying algorithm is prone to errors. Although the classification is generally accurate, single mistakes are common.
(d) Knowing that the acquired binary classifications are representing a continuous property (here: the original trajectory leading over or under the line in the middle), the initial classifications can be improved.

The corresponding transformations turning the observed view into a series of classifications are:

- **(a)**→(b): recording the video stream,
- **(b)**→(c): classifying the video stream as separate frames—the One Frame Analysis (OFA) approach,
- **(c)**→(d): improving the initial classifications—the proposed Frame Sequence Analysis (FSA) approach.

The visual distance between two pictures has to be defined in a domain-specific way. We will further on define it as a metric function $d(\cdot, \cdot)$, putting aside an exact derivation of its value. A real-life observation in a point of time t will be denoted as v_t. The first frame of the video starts at $t_0 = 0$, the time of frame number m ($m \in \mathbb{N}$) is therefore $t_m = m \cdot step$.

We observe the underlying continuous process (the real-life view) in discrete, evenly spaced moments (as video frames) and expect the amount of change to have an upper bound which allows for preserving the majority of the view from frame to frame. This kind of continuity is defined as

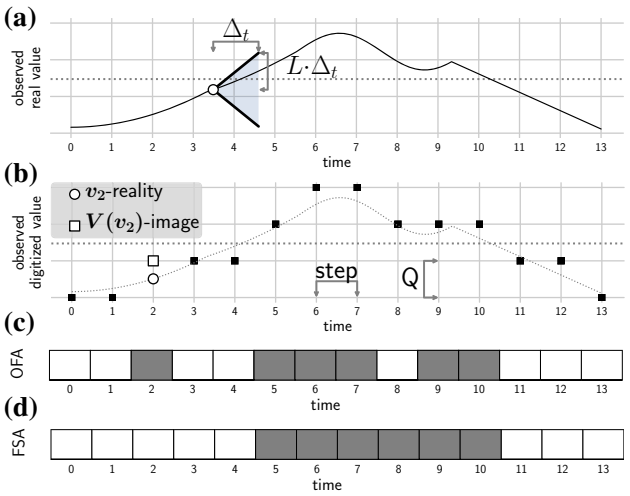

Fig. 1 Four stages from real-life view to final classification **a** Observed reality, **b** discrete and digitalized reality, where: step—time between two frames, Q—digitalization inaccuracy, i.e., pixels and discrete colors, **c** OFA result, **d** FSA result

Lipschitz continuity with the constant $L > 0$:

$$d(v_t, v_u) \leq L \cdot |t - u|. \tag{1}$$

This definition is illustrated in Fig. 1a. In a $\Delta_t = |t - u|$ time difference, the function can therefore change by no more than $\pm L \cdot \Delta_t$. For the exemplary point at $t = 3.5$, the constant L determines the limits within which the function can vary.

We interpret L as the maximal pace of change for the analyzed type of video. Assuming that the discrete sequence (frames and their classifications) is a view of a continuous function (i.e., the true view and state) which satisfies the Lipschitz property, the appropriate limitations on the change over time are still kept.

To show this, let us first define the function which represents the transformation of the real-life state into a discrete video. We will model it with the function $V(\cdot)$ which satisfies:

$$d(V(v_t), v_t) < \frac{Q}{2}. \tag{2}$$

The constant Q expresses the inaccuracy of the transformation and reflects the quality of the recording. The value of Q is lower for higher image quality.

Therefore, for two given observations v_s and v_t:

$$\begin{aligned} d(V(v_t), V(v_s)) &\leq d(V(v_t), v_t) + d(v_t, v_s) \\ &\quad + d(v_s, V(v_s)) \leq d(v_t, v_s) + Q. \end{aligned} \tag{3}$$

This result corresponds to the fact that the difference between two frames represents the difference between the views they represent and a limited inaccuracy of the recording (e.g., rounded colors, pixels).

Theorem 1 *A discrete view with a limited inaccuracy of a Lipschitz continuous function preserves the Lipschitz property.*

Proof Let us define: v_t—observation in point of time t, $V(\cdot)$—projection of the real-life view into a picture/frame, $p_m = V(v_{m \cdot step})$—discrete picture in the discrete moment $m \in \mathbb{N}$.
We will show that the Lipschitz property is still preserved for p_m. Let us take arbitrary different frame indexes m and m' ($\Delta_m = |m - m'| \geq 1$). Applying Eqs. 1 and 3, we get:

$$\begin{aligned} d(p_m, p_{m'}) &= d(V(v_{m \cdot step}), V(v_{m' \cdot step})) \\ &\leq d(v_{m \cdot step}, v_{m' \cdot step}) + Q \leq L \cdot step \cdot \Delta_m + Q \\ &\leq \left(L + \tfrac{Q}{step}\right) \cdot step \cdot \Delta_m \overset{Q' = \frac{Q}{step}}{=} (L + Q') \cdot step \cdot \Delta_m \end{aligned} \tag{4}$$

□

The proof shows that $(L + Q') \cdot step = L \cdot step + Q$ expresses the limit of change between consecutive frames. Thus, a faster pace of change can be compensated by lower *step* and Q values—which are limited only by the current technical development (highest possible frame rate and image quality). Already now, $Q \approx 0$, since the human eye often perceives videos as real-colored and cannot distinguish pixels.

Continuing, we will discuss the video frames in terms of their *ground truth* (GT) values, i.e., the actual value of the classified property. We will denote the ground truth of the frame m as G_m and its OFA classification as O_m ($G_m, O_m \in \{0, 1\}$). From the continuity of the analyzed video, we get that the closer two frames are in the video sequence, the bigger the chance that their classes in the GT are equal.

At this point, it is important to note the connection between Lipschitz continuity and the real-life origin of the analyzed videos. Observed real-life objects in a typical 25FPS video remain visible over multiple frames. A single positive outlier represents an error rather than an object (e.g., person) appearing for 0.04 s.

We will consider a frame's neighborhood as a time window of size w (odd, therefore $w = 2k + 1$ for a $k \in \mathbb{N}$). We adjust its size to be much smaller than the length $\mathcal{Z}$ of the current scene (sequence of consecutive positive or negative classifications in the GT). $\mathcal{Z}$ is a random variable with an unknown distribution. Due to the continuity of the video, we assume that its average value is significantly larger than 0. We choose a maximal window size $w_{\max}$ that ensures that the vast majority of windows are fully contained in a single scene:

$$P(w_{\max} \ll \mathcal{Z}) \approx 1. \tag{5}$$

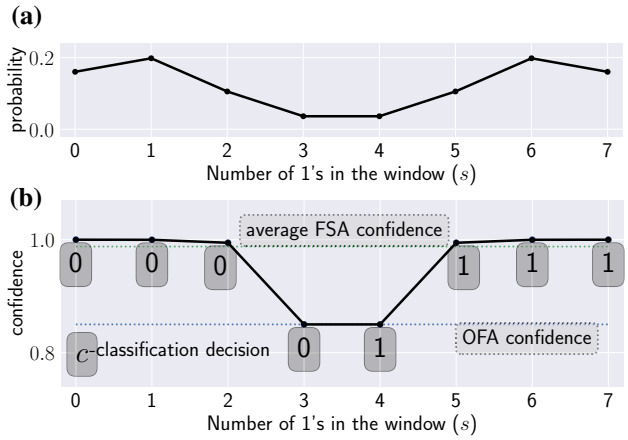

Fig. 2 Example calculations for Eqs. 7 and 8 ($w = 7$, $R_{1/1} = R_{0/0} = 0.85$): **a** distribution of windows with a given number of 1 s, **b** confidence of correct FSA corrections (black line) compared to the OFA confidence (0.85)

For two given values $c_1, c_2 \in \{0, 1\}$, we define the classification correspondence as:

$$R_{c_1/c_2} = P(O_m = c_1 \mid G_m = c_2). \tag{6}$$

These values can be estimated for a given OFA algorithm as the corresponding statistic of its performance on the GT data.

The probability of having a given number s of ones in a window of width $w = 2k+1$ comes from the observation that for both possible versions of the underlying GT the number of ones has a binomial distribution:

$$\begin{aligned} &P\left(\sum_{i=m-k}^{m+k} O_i = s\right) \\ &= \sum_{c\in\{0,1\}} P\left(\sum_{i=m-k}^{m+k} O_i = s | G_m = c\right) P(G_m = c) \\ &= \sum_{c\in\{0,1\}} \binom{s}{w} R_{1/c}^{s} \cdot R_{0/c}^{w-s} \cdot P(G_m = c). \end{aligned} \tag{7}$$

Continuing, we will establish the confidence of a decision indicated by a majority vote in the time window. It can be defined as the probability of the underlying GT being equal to 0 or 1 given that the window contains s positive classifications (using Bayes' theorem):

$$\begin{aligned} &P\left(G_m = c \mid \sum_{i=m-k}^{m+k} O_i = s\right) \\ &= \frac{P\left(\sum_{i=m-k}^{m+k} O_i = s \mid G_m = c\right) P(G_m = c)}{P\left(\sum_{i=m-k}^{m+k} O_i = s\right)}. \end{aligned} \tag{8}$$

The OFA algorithms improved by the proposed FSA method are expected to have a relatively high accuracy, since their output is the sole base for any further reasoning. Examples of applying Eqs. 7 and 8 are presented in Fig. 2. The numerical values provide a strong indication for leaning toward the majority result when deciding on the window center's classification assignment. The confidence of such a decision is at least at the level of OFA classifications.

FSA$_{w,\lambda,A}$ algorithm:
Input: Sequence of n OFA classifications $O[1 \ldots n]$
Output: Sequence of n FSA classifications $C[1 \ldots n]$
Parameters: w, λ, A
Algorithm:

1. **Init** C with O
2. **For** every window of width w in O:
3. Perform a weighted **vote** on the classification of the central frame with weights from the distribution D_λ
4. **If** the vote result exceeds the acceptance threshold A:
5. **Change** the central frame's classification in C
6. **return** C

Fig. 3 FSA pseudocode

In the next section, we will propose a new method based on the observations made above. The proposed method will be a more flexible, extended version of the majority vote considered so far, with two additional controlling parameters taken into account.

3 The FSA algorithm

The proposed approach corrects the results of a preliminary OFA classification. We have named it Frame Sequence Analysis (FSA). It is presented in Fig. 3. The controlling variables w, λ, A have been specified as parameters of the algorithm:

- w: window width, $5 \le w \le w_{\max}$;
- λ: distribution parameter, $0.2 < \lambda \le 1$;
- A: acceptance threshold, $0.5 \le A < 1$.

For any considered domain, the optimal values of those parameters may differ; therefore, they need to be established by means of discrete optimization. The minimal value $\lambda = 0.2$ has been set, because lower values of the parameter correspond to reducing the window size.

The significance distribution D_λ has been introduced to represent the decreasing relevance of frames further away from the window's center. It is a linear relation, adjusted by the parameter λ:

$$D_\lambda(i) = 1 - (1 - \lambda) \cdot \frac{|i|}{k} \quad \text{for } i \in \{-k, \ldots, k\}. \tag{9}$$

The weighted vote result for frame m is equal to:

$$\text{vote} = \frac{\sum_{j=-k}^{k} D_\lambda(j) O_{m+j}}{\sum_{j=-k}^{k} D_\lambda(j)} \tag{10}$$

(m is the index of the central frame in a shifting window, therefore: $k \leq m < n - k$). If the result of the vote exceeds the acceptance threshold A, then the result of the vote is deemed significant and its indicated value is assigned as the frame's classification (possibly changing the original value). Otherwise, the original OFA classification is kept. It is worth noting that by taking $A = 0.5$ and $\lambda = 1$ we acquire the majority voting variant discussed in the previous section.

4 Experiments

Classifications of images in videos are evaluated either in terms of frame-wise classification accuracy or scene boundary matching (with a certain lenience in terms of transitions between scenes). For our evaluation, we use four quality measures, describing different error rates:

- FNR: false negative ratio (computed $R_{0/1}$ value)
- FPR: false positive ratio (computed $R_{1/0}$ value)
- MBR_B: missing (scene) boundary ratio
- IBR_B: invalid (scene) boundary ratio,

where B is the number of frames by which a detected scene boundary's location can differ from its corresponding position in the GT. If a match is found within that distance, it is considered to be a correct detection. We define the values of MBR_B and IBR_B as follows:

$$\mathrm{MBR}_B = \frac{\#\,(\text{Missed Scene Boundaries})}{\#\,(\text{Scene boundaries in GT})}, \tag{11}$$

$$\mathrm{IBR}_B = \frac{\#\,(\text{Invalid Scene Boundaries})}{\#\,(\text{Detected scene boundaries})}. \tag{12}$$

An example of the evaluation of scene matching is presented in Fig. 4. The acquired error rates are $\mathrm{MBR}_2 = \frac{1}{2}$ and $\mathrm{IBR}_2 = \frac{3}{4}$.

The value of B expresses an acceptable discrepancy in scene boundary locations. Thus, it is considered to be domain specific. Consequently, there are four numerical values which vary between test executions:

- B: the scene boundary tolerance,
- w, λ, A: the controlling parameters of the FSA algorithm presented in Fig. 3.

Another variable of the experimental procedure is the test data. We have considered five sets of recordings:

- artificial stream with moving objects: slightly (AM1) or very (AM2) distorted,
- Chokepoint [22] first (CP1) and second (CP2) portal,
- Traffic Lights Recognition [23] (TLR).

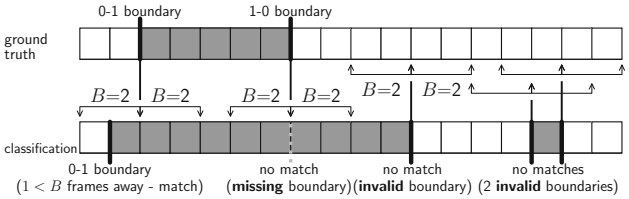

Fig. 4 Scene boundary tolerance example for $B = 2$

The artificial stream is a generated video sequence containing a drawing of a person and moving objects which overlap it. It is generated with a set of predefined distortions (blur, noise, random artifacts), to evaluate the influence of the video quality on OFA and FSA results. Chokepoint and TLR are open annotated datasets.

The characteristics of each of the datasets are different. The artificial stream contains regular movement of a limited number of objects and moderately long scenes. The Chokepoint recordings contain multiple short scenes, often with people passing through the view of the camera rapidly one after another. The TLR recording contains mostly very long scenes. Furthermore, the resolution of this video is the smallest.

Three OFA algorithms have been used as black-boxes, providing input for our FSA schemes:

- for the artificial streams - openCV [24] silhouette detection (with Haar cascades),
- for the Chokepoint datasets - openCV face detection (with Haar cascades),
- for Traffic Light Recognition - our implementation of the algorithm presented in [23].

The TLR dataset contains multiple kinds of annotations (green/yellow/red light or ambiguous). We have focused on detecting green lights and left out the ambiguous scenes, acquiring a binary OFA classifier.

The number of parameters allows us to perform a dense search of the parameter values' space. We consider best parameter values for every quality measure separately, as well as their combined root mean square.

The main point of interest of the testing procedure is the improvement in accuracy introduced by the FSA scheme, when it is compared with the underlying OFA algorithm. First, considering the structure of the results, we want to find relationships between parameters and provide guidelines for the final experimentation. We start with performing a simple exploratory analysis of the results acquired on the testing datasets. This is an introduction for evaluating the algorithm on the verification data.

The final experiment's scheme is presented in Fig. 5. To focus on the improvement our algorithm contributes to the considered OFA algorithms, the given FSA results are pre-

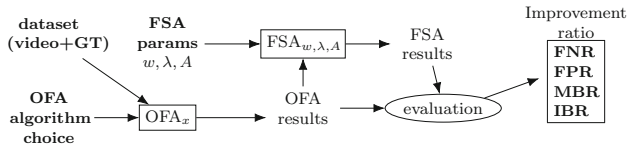

Fig. 5 Procedure for evaluating OFA and FSA classifications

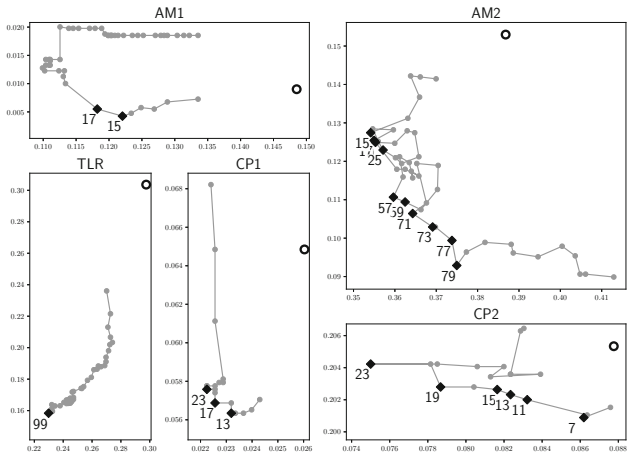

Fig. 6 FPR and FNR measures changing with window size (empty circle—OFA result)

sented as the ratio of every error type in the FSA output compared to the OFA output.

5 Experiment results

The first step of the experiments has been the analysis of the results acquired on training data. The ranges of possible parameter values have been densely covered with test executions. First, pairs of measures have been compared with each other for different values of w in a simple majority vote. Figure 6 shows the most definitive result of those comparisons (FPR/FNR). With an increasing w the measures improve steadily until a value, after which the quality of the FSA algorithm first stagnates and then changes chaotically. In contrast, for increasing values of w the IBR changed chaotically and MBR increased.

The values shown in Fig. 6 are two-dimensional; therefore, only a partial order can be introduced among them. The black markers represent the minimal results (i.e., there are no other window sizes acquiring a better improvement in terms of both measures at once), which are better than the original OFA. Those are compared with the typical sizes of scenes (ranges from the first to the third quartile) in the corresponding recordings in Table 1. The range of the best values of w correlates with the distribution of the scene size, but is also influenced by the quality of the data.

In Fig. 7, the best parameter values for each measure and some of their combinations have been presented. The results for A show that all measures besides MBR indicate a strong

Table 1 Best window sizes for scene lengths

Test case	AM1	AM2	TLR	CP1	CP2
Scene lengths	92–154	92–154	469–1425	57–79	53–80
Best window sizes	15–17	15–79	≥ 99	13–23	7–23

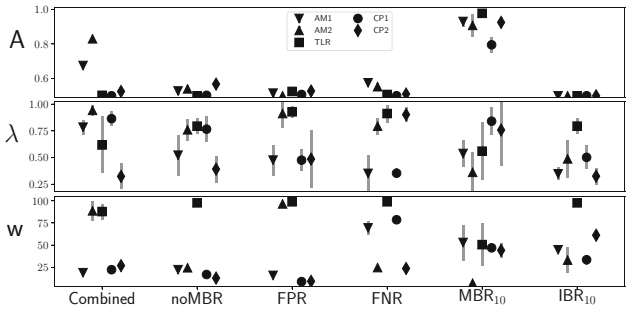

Fig. 7 Best parameter values for given measures. Mean and standard deviation range for best 25 results for each case

preference for lower values of A, which corresponds to their interpretations.

Intuitively, we expected the MBR measure to provide a soft limiting effect, so that the FSA step would not smoothen the results too much. It has been designed symmetrically to IBR, but has been shown to be much more unstable. In most tests, its value changed significantly, dominating the combined measure. The reason is as follows: The initial OFA classifications had been heavily segmented. Therefore, boundaries were rarely missing, which resulted in perfect MBR values in the initial data. Due to this strong influence, we conclude that its significance should either be limited or its value provided only as an informative measure after optimizing for the other measures.

To see the underlying relation between the optimal parameters, a plot of the best performing 100 parameter sets for each case is shown in Fig. 8.

Finally, parameter sets performing best on training data, have been evaluated on verification data and presented in Table 2. In virtually all cases the acquired error rates have been reduced. Figure 9 shows exemplary cases from CP2, where the FSA approach succeeds and fails in correcting classifications, what confirms two of our presumptions. First, the OFA detector makes short mistakes due to the variability of the appearance of detected objects in real-life videos—which is why it is susceptible to the FSA improvement. Secondly, the FSA method is not able to correct long-term errors, especially when the OFA algorithm fixates on generating false detections in a semi-static background.

Thus far, no universal methods have been proposed which could be compared with the FSA approach. Results in different works are comparable only to a limited degree, as evaluations are performed on different datasets, other underlying algorithms and with domain-specific measures. The

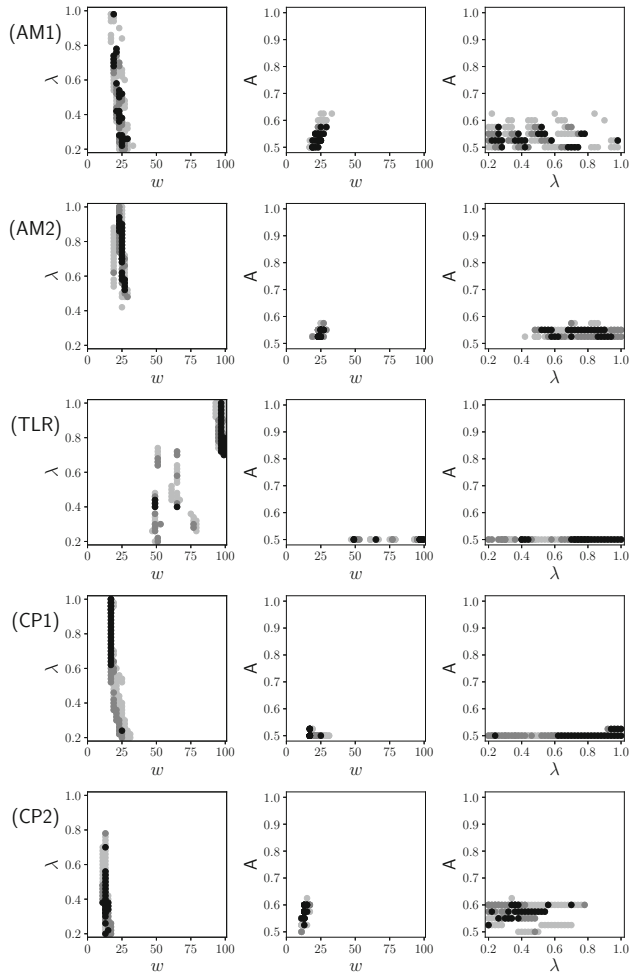

Fig. 8 Parameter relations ($B = 10$, best reduction of combined FPR, FNR and IBR; black—best result) for every dataset

temporal filtering in [10] improved a hand gesture recognition rate from 84.7 to 87.5%. The ratio of detections overlapping GT in [9] was improved 4% by a median temporal filter and 10% by a global HMM method. In [11], it was stated that outlier post-processing "slightly lowers" the FNR. In a different field (pedestrian reidentification [12]), but with an approach related to ours, an increase in the F-score values from 15.5 to 19.2% and from 25.6–28.1% to 33.5–38.9% has been acquired.

The results acquired by the FSA approach are of a similar magnitude to those acquired by these methods. This shows that the FSA step correctly and universally utilizes the temporal relations in the video. Also the fact that the FSA approach was able to improve all of the considered OFA algorithms (which are industry standard or state-of-the-art representatives) is a strong confirmation of the FSA approach and its universality.

Table 2 Quality results: FSA to OFA comparison

B	Dataset	FSA to OFA error ratio		
		Measure		
		FPR	FNR	IBR_B
5	AM1	0.92	0.89	0.76
	AM2	0.78	1.00	0.96
	TLR	0.97	1.02	1.00
	CP1	0.86	0.99	0.61
	CP2	0.95	0.98	0.65
10	AM1	0.83	0.88	0.75
	AM2	0.94	0.97	0.98
	TLR	0.56	0.87	0.93
	CP1	0.79	0.92	0.22
	CP2	0.96	1.00	0.58
20	AM1	0.83	0.90	0.35
	AM2	0.96	1.00	0.99
	TLR	0.56	0.87	0.91
	CP1	0.79	0.92	0.13
	CP2	0.95	0.99	0.24
50	AM1	0.71	0.91	0.35
	AM2	0.69	1.06	0.23
	TLR	0.56	0.87	0.88
	CP1	0.79	0.92	0.07
	CP2	0.95	0.99	0.16

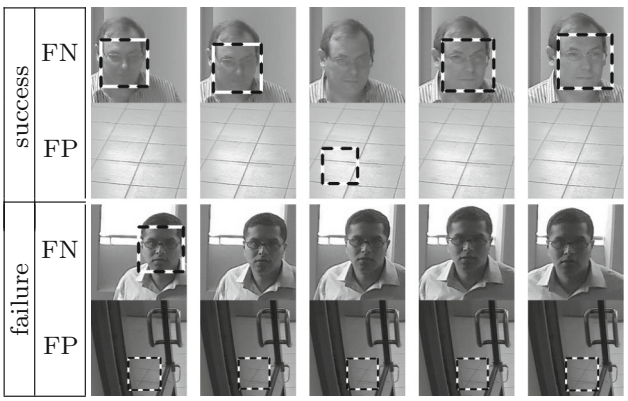

Fig. 9 Example OFA outputs: FSA success and failure cases

6 Conclusions

The proposed FSA method allowed to efficiently improve the considered algorithms, which classify images in video streams. It uses the temporal information in classification sequences, which in real-life videos are considered to represent underlying continuous processes. Therefore, the same approach can be applied to numerous other OFA algorithms.

We have presented a systematic and complete approach to defining and analyzing the properties of FSA schemes. After extending a simple majority voting scheme with additional

parameters (significance distribution D_λ and acceptance threshold A), we have shown how they influence all quality measures. In all cases, setting A close to 0.5 led to the best results. The best frame significance distribution parameter λ varied between cases, concentrating in different ranges of possible values. Other distributions can be analyzed, as our method allows to embed them. The relationship of the optimal window size w with the scene lengths' distribution fulfills the assumption from Eq. 5.

Our results have allowed to establish the applicability of the FSA method. It performs best in videos of non-perfect quality, where single OFA classifications are affected by distortions. In cases with a very high OFA accuracy level, the FSA step's effect would be smaller.

We propose continued experiments with different distributions and other parameterizations of the FSA scheme. Furthermore, an extended approach can be developed, which determines the significance distribution in the shifting window empirically, by measuring the similarity of frames. This could improve the FSA method quality, at the cost of computational efficiency.

We also propose to consider multi-categorical classifications, by using a binary encoding ("1 of n"—one-hot) of the category labels. A method of generalizing the FSA approach can be conceived, where the FSA step operates on single binary positions of the encoded classification before a label is chosen.

Finally, also further experimentation with new datasets and OFA algorithms is encouraged, to develop universal guidelines for the method's application.

Acknowledgements Research supported by *The Centre Of Competence For Novel Infrastructure Of Workable Applications* European Project POIG.02.03.00-22-059/13-00

References

1. Pan, G., Wang, L.: Swallowable wireless capsule endoscopy: progress and technical challenges. Gastroenterol. Res. Pract. **2012** (2012). https://doi.org/10.1155/2012/841691
2. Liu, Y., Zeng, L., Huang, Y.: An efficient HOG–ALBP feature for pedestrian detection. Signal Image Video Process. **8**(S1), 125–134 (2014)
3. Lan, X., Zhang, S., Yuen, P.C., Chellappa, R.: Learning common and feature-specific patterns: a novel multiple-sparse-representation-based tracker. IEEE Trans. Image Process. **27**(4), 2022–2037 (2018)
4. Lan, X., Ma, A.J., Yuen, P.C., Chellappa, R.: Joint sparse representation and robust feature-level fusion for multi-cue visual tracking. IEEE Trans. Image Process. **24**(12), 5826–5841 (2015)
5. Medouakh, S., Boumehraz, M., Terki, N.: Improved object tracking via joint color-LPQ texture histogram based mean shift algorithm. Signal Image Video Process. **12**(3), 583–590 (2018)
6. Froba, B., Kublbeck, C.: Face tracking by means of continuous detection. In: 2004 Conference on Computer Vision and Pattern Recognition Workshop, p. 65. IEEE (2004)
7. Lovell, B., Kootsookos, P.: Evaluation of HMM training algorithms for letter hand gesture recognition. In: Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No. 03EX795), pp. 648–651. IEEE (2003)
8. Huang, J., Liu, Z., Wang, Y.: Joint video scene segmentation and classification based on hidden Markov model. In: 2000 IEEE International Conference on Multimedia and Expo, vol. 3, pp. 1551–1554. IEEE (2000)
9. Liu, D., Chen, T.: Object Detection in Video with Graphical Models. In: 2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings, vol. 5, pp. 693–696. IEEE (2006)
10. Bourennane, S., Fossati, C.: Comparison of shape descriptors for hand posture recognition in video. Signal Image Video Process. **6**(1), 147–157 (2012)
11. Munzer, B., Schoeffmann, K., Boszormenyi, L.: Detection of Circular Content Area in Endoscopic Videos for Efficient Encoding and Improved Content Analysis. Technical Report, Institute of Information Technology, University Klagenfurt (2012)
12. Figueira, D., Taiana, M., Nascimento, J.C., Bernardino, A.: A window-based classifier for automatic video-based reidentification. IEEE Trans. Syst. Man Cybern. Syst. **46**(12), 1736–1747 (2016)
13. Gao, Z., Lu, G., Yan, P., Wang, L.: Retrospective analysis of time series for frame selection in surveillance video summarization. Signal Image Video Process. **11**(4), 581–588 (2017)
14. Haji-Maghsoudi, O., Talebpour, A., Soltanian-Zadeh, H., Haji-maghsoodi, N.: Automatic organs' detection in WCE. In: The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012), pp. 116–121. IEEE (2012)
15. Gupta, M., Gao, J., Aggarwal, C.C., Han, J.: Outlier detection for temporal data tutorial. In: 2013 SIAM International Conference on Data Mining. Austin, Texas, USA (2013)
16. Gupta, M., Gao, J., Aggarwal, C.C., Han, J.: Outlier detection for temporal data: a survey. IEEE Trans. Knowl. Data Eng. **26**(9), 2250–2267 (2014)
17. Basu, S., Meckesheimer, M.: Automatic outlier detection for time series: an application to sensor data. Knowl. Inf. Syst. **11**(2), 137–154 (2006)
18. Hill, D.J., Minsker, B.S.: Anomaly detection in streaming environmental sensor data: a data-driven modeling approach. Environ. Model. Softw. **25**(9), 1014–1022 (2010)
19. Cao, Y., Liu, D., Tavanapong, W., Wong, J., Oh, J., De Groen, P.C.: Computer-aided detection of diagnostic and therapeutic operations in colonoscopy videos. IEEE Transa. Biomed. Eng. **54**(7), 1268–1279 (2007)
20. Duchenne, O., Laptev, I., Sivic, J., Bach, F., Ponce, J.: Automatic annotation of human actions in video. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1491–1498. IEEE (2009)
21. Sun, Z., Hu, Z., Chiong, R., Wang, M., Zhao, S.: An adaptive weighted fusion model with two subspaces for facial expression recognition. Signal Image Video Process. **12**(5), 835–843 (2018)
22. Wong, Y., Chen, S., Mau, S., Sanderson, C., Lovell, B.C.: Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In: CVPR 2011 Workshops, pp. 74–81. IEEE (2011)
23. Charette, R.D., Nashashibi, F.: Real time visual traffic lights recognition with image processing. Adv. Robot. **33**, 358–363 (2009)
24. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)