This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:10.1109/TCSVT.2015.2392531, IEEE Transactions on Circuits and Systems for Video Technology

Real-time Fire Detection for Video Surveillance Applications using a Combination of Experts based on Color, Shape and Motion

Pasquale Foggia, Alessia Saggese and Mario Vento, IAPR Fellow
Dept. of Information Eng., Electrical Eng. and Applied Mathematics
University of Salerno, Italy
{pfoggia, asaggese, mvento}@unisa.it

***Abstract*—In this paper we propose a method able to detect fires by analyzing the videos acquired by surveillance cameras. Two main novelties have been introduced: first, complementary information, respectively based on color, shape variation and motion analysis, are combined by a multi expert system. The main advantage deriving from this approach lies in the fact that the overall performance of the system significantly increases with a relatively small effort made by designer. Second, a novel descriptor based on a bag-of-words approach has been proposed for representing motion. The proposed method has been tested on a very large dataset of fire videos acquired both in real environments and from the web. The obtained results confirm a consistent reduction in the number of false positives, without paying in terms of accuracy or renouncing the possibility to run the system on embedded platforms.**

***Index Terms*—Video surveillance, Fire Detection, Multi Expert System**

I. Introduction

In the last years several methods have been proposed, with the aim to analyze the videos acquired by traditional video-surveillance cameras and detect fires or smoke, and the current scientific effort [1] [2] focused on improving the robustness and performance of the proposed approaches, so as to make possible a commercial exploitation.

Although a strict classification of the methods is not simple, two main classes can be distinguished, depending on the analyzed features: color based and motion based. The methods using the first kind of features are based on the consideration that a flame, under the assumption that it is generated by common combustibles as wood, plastic, paper or other, can be reliably characterized by its color, so that the evaluation of the color components (in RGB, YUV or any other color space) is adequately robust to identify the presence of flames. This simple idea inspires several recent methods: for instance, in [3] and [4] fire pixels are recognized by an advanced background subtraction technique and a statistical RGB color model: a set of images have been used and a region of the color space has been experimentally identified, so that if a pixel belongs to this particular region, then it can be classified as fire. The main advantage of such algorithms lies in the low computational cost allowing the processing of more than 30 frames per second at QCIF (176x144) image resolution. Differently from [3] and [4], in [5] the authors experimentally define a set of rules for filtering fire pixels in the HSI color space. The introduction of the HSI color space significantly simplifies the definition of the rules for the designer, being more suitable for providing a people-oriented way of describing the color. A similar approach has been used in [6], where a cumulative fire matrix has been defined by combining RGB color and HSV saturation: in particular, starting from the assumption that the green component of the fire pixels has a wide range of changes if compared with red and blue ones, this method evaluates the spatial color variation in pixel values in order to distinguish non-fire moving objects from uncontrolled fires.

The common limitation of the above mentioned approaches is that they are particularly sensitive to changes in brightness, so causing a high number of false positive due to the presence of shadows or to different tonalities of the red. This problem can be mitigated by switching to a YUV color space. In [7], for instance, a set of rules in the YUV space has been experimentally defined to separate the luminance from the chrominance more effectively than in RGB, so reducing the number of false positives detected by the system. In [8] information coming from YUV color are combined using a fuzzy logic approach in order to take into account the implicit uncertainties of the rules introduced for thresholding the image. A probabilistic approach based on YUV has been also exploited in [9], where the thresholding of potential fire pixels is not based on a simple heuristic but instead on a support vector machine, able to provide a good generalization without requiring problem domain knowledge. Although this algorithm is less sensitive to variations in the luminance of the environment, its main drawback if compared with other color based approaches lies in the high computational cost required as soon as the dimensions of the support vector increase.

In conclusion, it can be observed that the methods using color information, although being intrinsically simple to configure, can be successfully used only in sterile areas, where no objects generally move inside. In fact, their main limitation concerns the number of false positives when used in normal populated areas: persons with red clothes or red vehicles might be wrongly detected as fire only because of their dominant color. In order to face this issue, in the last years several approaches have been proposed: they start from the assumption that a flame continuously changes its shape and the disordered movement of red colored regions can help in distinguishing it

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:10.1109/TCSVT.2015.2392531, IEEE Transactions on Circuits and Systems for Video Technology

from rigid objects moving in the scene. For instance, in [10] the physical properties of the fire are used to build a feature vector based on an enhanced optical flow, able to analyze in different ways both the dynamic texture of the fire and its saturated flame. Dynamic textures have also been used in [11], where a two-phase texture detection process has been proposed in order to speed-up the segmentation step, very useful to extract a wide set of shape-based features, and making possible the detection of the fire in a reasonable time. In [12] the irregularity of the fire over time is handled by combining the capabilities of finite state automata with fuzzy logic: variations in wavelet energy, motion orientation and intensity are used to generate probability density functions, which determine the state transitions of a fuzzy finite automaton.

The Wavelet transform has been also used in [13] in order to properly detect the temporal behavior of flame boundaries. It is worth pointing out that the methods based on the wavelet transform, differently from those based on the color, cannot be used on still images and in general require a frame rate sufficiently high, higher than 20 fps, to guarantee satisfactory results, so limiting their applicability.

In [14] frame-to-frame changes are analyzed and the evolution of a set of features based on color, area size, surface coarseness, boundary roughness and skewness is evaluated by a Bayesian classifier. The wide set of considered features allows the system to take into account several aspects of fire, related to both color and appearance variation, so increasing the reliability in the detection. In [15] the thresholding on the color, performed in the RGB space, is improved by a multi resolution two-dimensional wavelet analysis, which evaluates both the energy and the shape variations in order to further decrease the number of false positive events. In particular, the shape variation is computed by evaluating the ratio between the perimeter and the area of the minimum bounding box enclosing the candidates fire pixels. This last strategy is as simple and intuitive as promising if the scene is populated by rigid objects, such as vehicles. On the other side, it is worth pointing out that the shape associated to non rigid objects, such as people, is highly variable in consecutive frames: think, for instance, to the human arms, that may contribute to significantly modify the size of the minimum bounding box enclosing the whole person. This evidence implies that the disordered shape of the person may be confused with the disordered shape of the fire, so consistently increasing the number of false positives detected by the system.

So, in conclusion, the main limitation of motion based approaches lies in the fact that the performance improvement is often paid from different points of view: first, in most of the cases, several sensitive parameters need to be properly set for the application at hand. Second, the motion and the shape of the flame is somehow dependent on the burning material, as well as on the weather conditions (think, as an example, of a strong wind moving the fire).

In [16] a novel descriptor based on spatio-temporal properties is introduced. First, a set of 3D blocks is built by dividing the image into 16×16 squares and considering each square for a number of frames corresponding to the frame rate. The blocks are quickly filtered using a simple color model of the flame pixels. Then, on the remaining blocks a feature vector is computed using the covariance matrix of 10 properties related to color and to spatial and temporal derivatives of the intensity. Finally, an SVM classifier is applied to these vectors to distinguish fire from non-fire blocks. The main advantage deriving from this choice is that the method does not require background subtraction, and thus can be applied also to moving cameras. However, since the motion information is only taken into account by considering the temporal derivatives of pixels, without an estimation of the motion direction, the system, when working in non sterile areas, may generate false positives due, for instance, to flashing red lights.

The idea of combining several classifiers to obtain a more reliable decision has been generalized and extended in a theoretically sounder way in the pioneering paper [17]. Fire colored pixels are identified by using a Hidden Markov Model (HMM); temporal wavelet analysis is used for detecting the pixel flicker; spatial wavelet analysis is used for the non-uniform texture of flames; finally, wavelet analysis of the object contours is used to detect the irregular shape of the fire. The decisions taken by the above mentioned algorithms are linearly combined by a set of weights which are updated with a least-mean-square (LMS) strategy each time a ground-truth value is available. This method has the advantage that during its operation, it can exploit occasional feedback from the user to improve the weights of the combination function. However, a drawback is the need to properly choose the learning rate parameter, in order to ensure that the update of the weights converges, and that it does so in a reasonable time.

The paper is organized as follows; Section II describes the rationale of our approach; in Section III the proposed method is detailed: after a description of the MES in Subsection III-A, the three different experts, based on color, shape variation and motion, are described in Subsections III-B, III-C and III-D, respectively. In Section IV the results obtained by testing the proposed approach over a wide dataset are shown, before drawing some conclusions in Section V.

II. Rationale of the method

Up to now the efforts of the research community have been mainly devoted to the definition of a representation of both color and movement, so as to discriminate fire from non fire objects; this inevitably leads to high dimensional feature vectors. How to manage high dimensional feature vectors is a well-known problem in the communities of machine learning and pattern recognition: in fact, as shown in [18], employing a high dimensional feature vector would imply a significant increase in the amount of data required to train any classifier in order to avoid overspecialization and to achieve good results. Furthermore, independently of the particular features extracted, in most of the above mentioned methods the high variability of fires, as well as the large amount of noise in data acquired in fire environments, prevent the systems from the achievement of a high recognition rate. More generally, it has been shown [19] that increasing the performance of a system based on the traditional combination feature vector – classifier is often a very expensive operation: in fact, it may

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:10.1109/TCSVT.2015.2392531, IEEE Transactions on Circuits and Systems for Video Technology

require to design a new set of features to represent the objects, to train again the classifier, or to select a different classifier if the performance are not sufficiently satisfactory. Furthermore, this effort could be payed back by only a slight improvement in the overall accuracy, so this approach may prove not very convenient.

In order to overcome the above mentioned limitations, one of the solutions coming from the literature [19] is to split the feature vector and consequently to adopt a set of classifiers, each tailored on a feature set and then trained to be an expert in a part of the feature space. The main idea of this kind of paradigm, usually referred to as Multi Expert System (MES), is to make the decision by combining the opinions of the different individual classifiers (hereinafter *experts*), so as to consistently outperform the single best classifier [20]. This latter paper explains on the basis of a theoretical framework why a MES can be expected to outperform a single, monolithic classifier. In fact, most classifiers, given an unlimited amount of training data, converge to an optimal classification decision (in a probabilistic sense); but on a finite training set, their output is affected by an error (additional with respect to the inherent error due to ambiguities in the input data), which is either due to overspecialization or to the choice of reducing the classifier complexity in order to avoid the loss of generalization ability. The author of [20] shows that, under some assumptions satisfied very often in practical cases, a suitably chosen *benevolent* combining function can make the overall output of the MES less sensitive to the errors of the individual classifiers.

MESs have been successfully applied in several application domains, ranging from biomedical images analysis [21] [22] and face detection [23] to movie segmentation [24] and hand-writing recognition [19]. Starting from a preliminary study [25], in this paper we propose the employing of a MES for detecting the fire in both indoor and outdoor environments: three different experts, complementary in their nature and regarding their errors, are combined with relatively little effort so as to make possible the improvement of the overall performance. It is evident that the successful implementation of a MES requires both the adoption of complementary sets of features feeding the different experts and the choice of a reliable combination rule.

Regarding the first aspect, we considered three different experts able to analyze the same problem from different points of view, based on color, on movement and on shape variation respectively. The main advantage deriving from this choice lies in the fact that the experts are very simple to configure, so making the proposed system particularly suited for deployment in real environments.

As for the experts based on color and shape variation, two algorithms widely adopted by the scientific community and providing very promising results have been considered; they are based on a thresholding in the YUV space and on the variation of the shape in terms of minimum bounding box enclosing the detected moving object, respectively. In particular, the expert based on color aims to discriminate *red* from *non red* objects and is particularly suited for sterile environments, while the one based on shape variation is very effective for distinguishing fire, usually having a strongly variable shape, from rigid objects moving in the scene, such as vehicles.

Finally, the expert based on movement evaluation is based on the assumption that fire has a disordered movement, much more disordered if compared with any other object usually populating the scene. In order to exploit this property, a novel descriptor for representing the movement is proposed in this paper: the main idea is to adopt a bag of words approach for evaluating the direction of some salient points detected in the moving objects. The main advantage deriving from this choice is that the representation is very robust with respect to the noise introduced both during the extraction of the salient points and the evaluation of the direction of their motion.

Once obtained the decisions from the three different experts, the system needs to properly combine them: the idea is that each classifier should have different voting priorities, depending on its own learning space. For this reason, the combination rule adopted in this paper is based on weighted voting, where the weights depend on the prediction confidence of each class the system has to recognize [20].

The main original contributions of the paper are: 1) the proposition of a novel system for characterizing the movement of the flame; 2) the use of a multi expert approach based on three complementary experts; 3) a wide characterization of performance on a standard dataset of videos, made available at *http://mivia.unisa.it*.

III. Proposed Architecture

An overview of the proposed approach is presented in Figure 1. Objects moving in the scene are first detected by using the algorithm we recently proposed in [26], which proved to be very effective both from a qualitative and a computational point of view: a model of the background (which represents the scene without any object moving inside) is maintained and properly updated (*Background updating*) so as to deal with the changes of the environments during the day; then, a background subtraction strategy is applied in order to obtain the foreground mask, encoding the objects moving in the scene (*Foreground mask extraction*). Finally, the *blobs*, each one being associated to an object, are obtained by a connected component labeling analysis [27] (*Connected component labeling*).

Three different experts have been introduced for evaluating the blobs: the first is based on color (*Color evaluation*, hereinafter *CE*); the second analyzes the shape of the blobs detected in the current frame with respect to the ones detected in previous frame (*Shape variation*, hereinafter *SV*); the third evaluates the movement of the blobs in two consecutive frames (*Movement evaluation*, hereinafter *ME*). The decisions taken by the experts are combined by a MES classifier based on a weighted voting rule, which finally assigns a class to each blob.

A. Multi expert evaluation

As mentioned before, one of the main choices determining the performance of a MES is the combination rule. Although

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:10.1109/TCSVT.2015.2392531, IEEE Transactions on Circuits and Systems for Video Technology

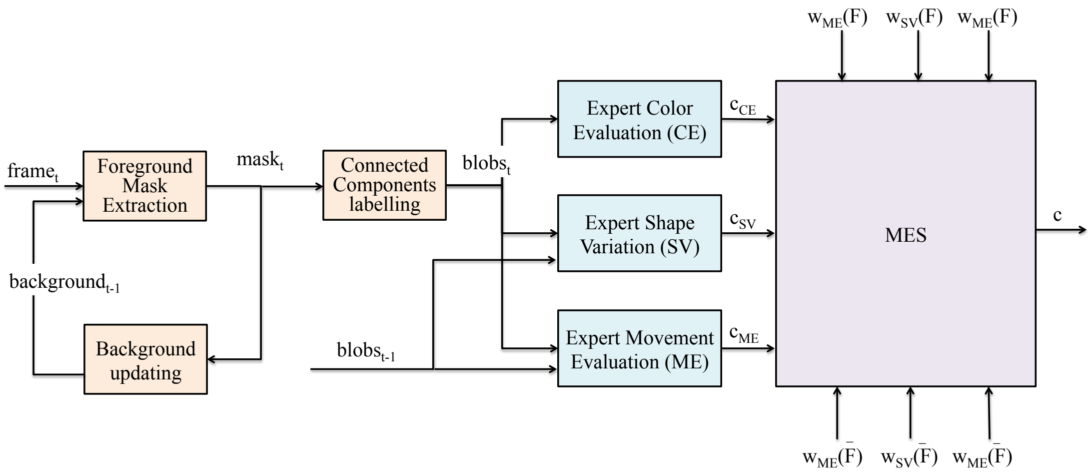


Fig. 1: Overview of the proposed approach: the blobs are detected by a background subtraction algorithm; the decision is taken by combining information coming from three different experts, respectively based on color, shape variation and motion which analyze every input blob. Note that the last two experts need to be supplied with the blobs detected at the current frame and at the previous frame.

several strategies have been proposed in the last years [28], it has been proved that one of the most robust to errors of the combined classifiers (both when combining classifiers based on the same features and when using different feature subsets in each classifier) is the weighted voting rule [20]. The main idea is that each expert can express its vote, which is weighted proportionally to the recognition rate it achieved for each class on the training set. For instance, let suppose that both CE and the ME classify the blob b as fire and that the percentage of fires correctly detected on the training set is 0.8 and 0.7 for the two experts respectively. Then, the two experts' votes for the class fire will be weighted 0.8 and 0.7.

In a more formal way, the generic k-th expert, being $k \in \{CE, SV, ME\}$, assigns to the input blob the class $c_k(b)$ chosen between the labels (F for *fire*, $\overline{F}$ for *non fire*); this can be formulated as a vote to the generic class i as follows:

$$\delta_{ik}(b) = \begin{cases} 1 & \text{if } c_k(b) \text{ gives the class } i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In other words, if the output corresponds to the class, then the vote will be 1, otherwise it will be 0.

As suggested in [19], the weights $w_k(i)$ are dynamically evaluated by a Bayesian formulation in order to lead to the MES highest recognition rate. In particular, this formulation takes into account the performance of each expert on the training set of each class. More formally, given the classification matrix $C^{(k)}$ computed by the k-th expert on the training step, $w_k(i)$ can be determined by evaluating the probability that the blob b under test, belonging to the class i, is assigned to the right class c_k by the k-th expert:

$$w_k(i) = P(b \in i | c_k(b) = i) = \frac{C_{ii}^{(k)}}{\sum_{i=1}^{M} C_{ij}^{(k)}}, \tag{2}$$

being M the number of classes and $C_{(ij)}$ the value of the classification matrix in the position (i, j).

The final decision is taken by maximizing the reliability of the whole MES in recognizing that particular class. In particular, the reliability $\psi(i)$ that the blob b belongs to the class i is computed by a weighted sum of the votes:

$$\psi(i) = \frac{\sum_{k \in \{CE,SV,ME\}} \delta_{ik}(b) \cdot w_k(i)}{\sum_{k \in \{CE,SV,ME\}} w_k(i)}. \tag{3}$$

The decision for the class c is finally taken by maximizing the reliability over the different classes:

$$c = \arg\max_i \psi(i). \tag{4}$$

B. CE: the expert based on color evaluation

This expert evaluates the color by analyzing its properties in the YUV color space; as already mentioned, YUV has been widely adopted in the literature since it separates the luminance from the chrominance and so is less sensitive to changes in brightness. In particular, as proposed on [7], this expert is based on the combination of six different rules, denoted as $r_1^c \dots r_6^c$, able to model the color of the flames.

In more details, as for r_1^c and r_2^c, the idea is related to the experimental evidence that in most of flames the pixels exhibit a Red channel value greater than Green, as well as a Green greater than Blue [29]:

$$R(x,y) > G(x,y) > B(x,y). \tag{5}$$

Such conditions can be equivalently expressed in the YUV plane, by adopting the well known conversion rules [30], so that we obtain for the generic pixel (x,y) of the image:

$$r_1^c : Y(x,y) > U(x,y); \tag{6}$$

$$r_2^c : V(x,y) > U(x,y) \tag{7}$$

On the other side, r_3^c and r_4^c are based on the assumption that the Red component of fire pixels is higher than the mean Red component in the frame. Expressed in the YUV space, it

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:10.1109/TCSVT.2015.2392531, IEEE Transactions on Circuits and Systems for Video Technology

implies that a fire pixel has the Y and V components higher than the mean Y and V value in the frame respectively, while the U component lower than the mean U value in the frame:

$$r_3^c : Y(x,y) > \frac{1}{N} \cdot \sum_{k=1}^{N} Y(x_k, y_k) \tag{8}$$

$$r_4^c : U(x,y) < \frac{1}{N} \cdot \sum_{k=1}^{N} U(x_k, y_k), \tag{9}$$

$$r_5^c : V(x,y) > \frac{1}{N} \cdot \sum_{k=1}^{N} V(x_k, y_k), \tag{10}$$

being N the total number of pixels in the image.

Finally, in [7] it has been experimentally evaluated that fire pixels are characterized by a considerable difference between U and V components. Thus, the last rule can be defined as:

$$r_6^c : |V(x,y) - U(x,y)| \geq \tau_c. \tag{11}$$

In our experiments, τ_c has been set to 40, as suggested in [7].

The classifier decision c_{CE} is finally taken by evaluating the above mentioned rules. In particular, if all the conditions are verified, then the blob is assigned to the fire class:

$$c_{CE} = \begin{cases} F & \text{if } r_1^c \wedge r_2^c \wedge r_3^c \wedge r_4^c \wedge r_5^c \wedge r_6^c \\ \overline{F} & \text{otherwise} \end{cases} \tag{12}$$

C. SV: the expert based on shape variation

This expert (hereinafter referred as *SV*) analyzes the variation of the blob shape across two consecutive frames in order to exploit the observation that the shape of flames changes very quickly. In particular, as in [15], the algorithm computes, for each blob, the perimeter P_t and the area A_t of the minimum bounding box enclosed it. Such values are then used to compute the perimeter-area ratio r_t, which is an indicator of shape complexity:

$$r_t = \frac{P_t}{A_t}. \tag{13}$$

The shape variation s_v^t is then evaluated by comparing the shape measure computed at the frame t with the one obtained by the nearest blob detected at the previous frame $(t-1)$:

$$s_v^t = \left| \frac{r_t - r_{t-1}}{r_t} \right|. \tag{14}$$

The score s_v^t is finally analyzed; if it is higher than a given threshold, then the class fire is assigned to the input blob:

$$c_{SV} = \begin{cases} F & \text{if } s_v^t > \tau_v \\ \overline{F} & \text{otherwise} \end{cases} \tag{15}$$

D. ME: the expert based on movement evaluation

ME is based on a novel descriptor which adopts a bag-of-words approach [31], introduced in this paper in order to characterize the cluttered movement of fire. The rationale of this expert is based on the empirical observation that the parts of a flame appear to move at the same time in several different directions in a rather chaotic and unpredictable way; by contrast the parts of a rigid or articulated object show at each instant a quite limited set of motion directions. For translating this observation into an effective description and classification system, we have chosen a bag-of-words approach.

Bag-of-words has been successfully applied in several application domains, ranging from text classification to audio event detection and action recognition. The underlying idea is that the pattern to be classified is represented by the occurrences of low-level features (words) belonging to a dictionary and such occurrences are used to build a high-level vector; the generic component is associated to a word and its value is given counting to the occurrences of that word in the input pattern (see Figure 2 for an example).

In order to apply the bag of words strategy to our problem, the following main steps need to be dealt with: the extraction of the low-level representation, the definition of the dictionary that determines the construction of the high-level representation, and the paradigm adopted for the classification. An overview of the proposed approach is shown in Figure 4, while a more detailed explanation of the above mentioned phases will be provided in the following.

Low-level representation In order to capture the motion of the different parts of a foreground blob, salient points are extracted and matched across consecutive video frames. The set of salient points is extracted by using the Scale Invariant Feature Transform (SIFT) [32] descriptors. At a given time instant t, the high-speed corner detection algorithm proposed in [33] is used to extract the set C_t of $|C_t|$ corners:

$$C_t = \{c_t^1, ..., c_t^{|C_t|}\}. \tag{16}$$

Each corner is then represented by measuring the local image gradients in the region around it, so obtaining the set of corresponding SIFT descriptors:

$$V_t = \{v_t^1, ..., v_t^{|V_t|}\}, \tag{17}$$

being $|V_t| = |C_t|$.

Given the corner points extracted in two consecutive frames (t and $t-1$) and the corresponding set of descriptors (V_t and V_{t-1}), the algorithm computes a set of matchings by pairing each point at time t with the one at time $t-1$ that is closest according to the Euclidean distance between SIFT descriptors:

$$M = \{m_1 \ldots m_{|M|}\} \tag{18}$$

where:

$$m_j = (v_t^a, v_{t-1}^b) \tag{19}$$

such that:

$$b = \arg\min_i \|v_t^a - v_{t-1}^i\| \text{ and } \|v_t^a - v_{t-1}^b\| < \tau_M \tag{20}$$

For each matching m_j we consider the vector connecting the two corresponding corner points c_t^a and c_{t-1}^b, and extract the angle ϕ_j of this vector with respect to the x axis.

Figure 4 clarifies this concept: the corner points c_{t-1} and c_t, represented as red and blue circles respectively, are associated to their descriptors v_{t-1} and v_t. The matching m_j is represented by the green line connecting such points, while ϕ_j is the angle that m_j build with the x axis (black line).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:10.1109/TCSVT.2015.2392531, IEEE Transactions on Circuits and Systems for Video Technology

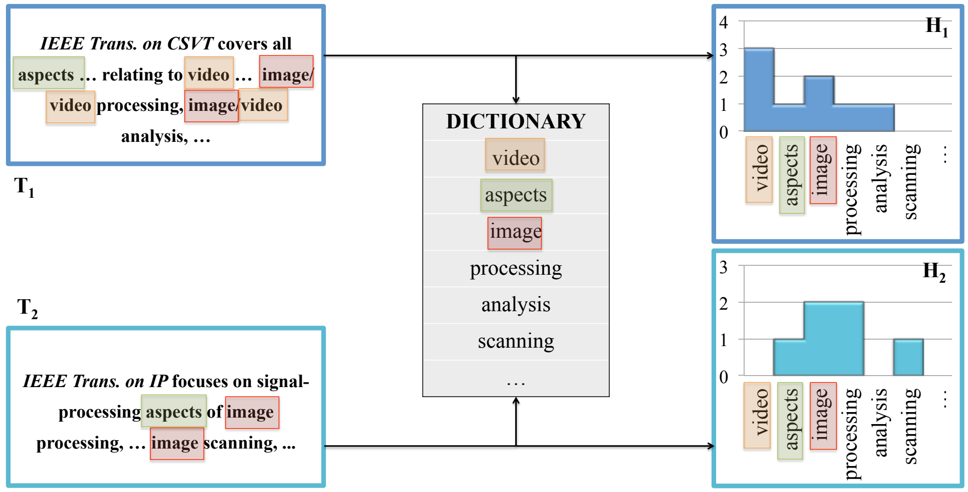


Fig. 2: The rationale of the Bag of words approach applied to text classification: the occurrences, in a document, of a predefined set of words included into a dictionary are used for building up an histogram of occurrences. See the histograms H_1 and H_2 associated to texts T_1 and T_2 respectively.

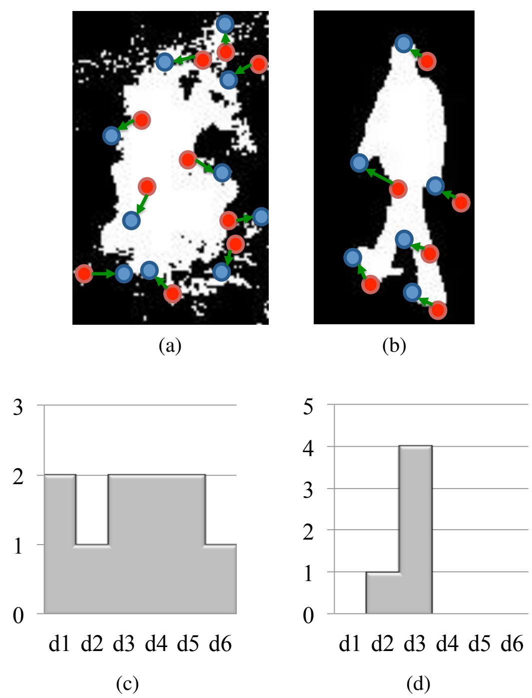


Fig. 3: Low level representation (a,b) and high level representation (c,d) of a fire (a,c) and a non fire (b,d) blob. Red and blue circles in (a) and (b) represent the salient points extracted at the frame t and $t-1$ respectively.

Dictionary According to the proposed low-level representation, the space of the possible words is the round angle ($0^\circ - 360^\circ$). In order to obtain a adequately small finite set of words, we decide to uniformly partition the space into $|D|$ sectors d, so obtaining the dictionary D as follows:

$$D = \left\{ d_k \middle| \quad k = 1, \ldots, |D|; \quad d_k = \left] k\frac{2\pi}{|D|}, (k+1)\frac{2\pi}{|D|} \right] \right\}. \quad (21)$$

$|D|$ has been experimentally set in this paper to 6: it implies that the resolution of the direction is 60°, which represents a good tradeoff between a suitable representation of the movement and the immunity to the noise.

High-level representation Given the dictionary D, for each blob the algorithm finds the words of D that occur in the blob, i.e. the intervals d_k that correspond to the motion direction of the salient points; then the blob can be represented by the histogram H of the occurrences of such words. An example is reported in Figure 3, where the low level representation (a,b) and the corresponding high level representation (c,d) for a fire (a,c) and a non fire object (b,d) is shown.

In a more formal way, the generic angle ϕ_j is associated to the index s_j, depending on the word d_k it belongs to:

$$s_j = |k : \phi_j \in d_k, k \in \{1, \ldots, |D|\}|. \quad (22)$$

The set $S = \{s_1, \ldots, s_{|M|}\}$ associated to a generic blob b is then evaluated obtaining the histogram $H = \{h_1, \ldots, h_{|D|}\}$, whose generic element h_i can be computed as follows:

$$h_i = \sum_{l=1}^{|M|} \delta(s_l, i), \quad i = 1, \ldots, |D|, \quad (23)$$

being $\delta(\cdot)$ the Kronecker delta.

Classification The main assumption used for designing the classifier is the evidence that the obtained feature vector is different for the two classes; in presence of fire the movement is disordered, determining occurrences of the words rather homogeneously distributed. Vice versa, when a rigid or articulated object moves in the scene, we mainly obtain values concentrated on a single or a few bins. See Figure 3 for an example.

For this reason, we introduce a measure of the homogeneity hm of the histogram:

$$hm = 1 - \frac{max(H)}{\sum_{k=1}^{|H|} h_k} \quad (24)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:10.1109/TCSVT.2015.2392531, IEEE Transactions on Circuits and Systems for Video Technology

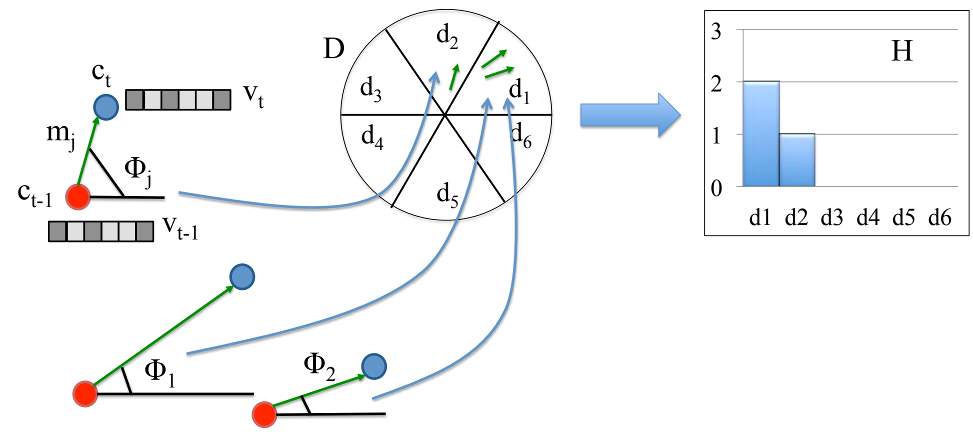


Fig. 4: Given the corner points c_{t-1} and c_t (red and blue circles respectively), the matching m_j is obtained by minimizing the euclidean distance between the corresponding descriptors v_{t-1} and v_t. The direction of the motion, encoded by the angle ϕ_j, is evaluated according the dictionary D and the histogram of occurrences H is then built.

and consequently if it is higher than a given threshold, then the input is classified as fire, otherwise as not fire:

$$c_{ME} = \begin{cases} F & \text{if } hm > \tau_m \\ \overline{F} & \text{otherwise} \end{cases} \quad (25)$$

IV. Experimental results

Most of the the methods in the literature (especially the ones based on the color evaluation) are tested using still images instead of videos. Furthermore, no standard datasets for benchmarking purposes have been made available up to now. One of the biggest collection of videos for fire and smoke detection has been made available by the research group of Cetin [34] [13]. Starting from this collection, composed by approximatively 31.250 frames, we added several long videos acquired in both indoor and outdoor situations so resulting in a new dataset composed by 62.690 frames and more than one hour of recording. More information about the different videos are reported in Table I, while some visual examples are shown in Figure 5[1].

Note that the dataset can be seen as composed by two main parts: the first 14 videos characterized by the presence of fire and the last 17 videos which do not contain fires; in particular, this second part is characterized by objects or situations which can be wrongly classified as containing fire: a scene containing red objects may be misclassified by color based approaches, while a mountain with smoke, fog or clouds may be misclassified by motion based approaches.

Such composition allows us to stress the system and to test it in several conditions which may happen in real environments.

The dataset has been partitioned into two parts: 80% has been used to test the proposed approach while 20% for training the system by determining the weights of the MES.

An overview of the performance achieved on the test set, both in terms of accuracy and false positives, is summarized in Table II.

[1]The whole dataset can be downloaded from our website: *http://mivia.unisa.it/datasets/video-analysis-datasets/fire-detection-dataset/*.

Among the three experts considered in this paper (CE, ME and SV), the best one is the CE, which achieves on the considered dataset a very promising performance (accuracy = 83.87% and false positives = 29.41%). Note that such performance is comparable with the one reached by the authors in [7], where over a different dataset the number of false positives is about 31%.

On the other hand, we can also note that the expert ME, introduced for the first time in this paper for identifying the disordered movement of fire, reveals to be very effective. In fact, we obtain a 71.43% accuracy and 53.33% false positives. It is worth pointing out that the considered dataset is very challenging for this expert: in fact, the disordered movement of smoke as well as of trees moving in the forests can be easily confused with the disordered movement of the fire. This consideration explains the high number of false positives introduced by using only ME.

As expected, the best results are achieved by the proposed MES, which outperforms all the other methods, both in terms of accuracy (93.55%) and false positives (11.76%). The very low false positive rate, if compared with state of the art methods, is mainly due to the fact that ME and SV act, in a sense, as a filter with respect to CE. In other words, ME and SV are able to reduce the number of false positives introduced by CE without paying in terms of accuracy: this consideration is confirmed by the results shown in Figure 7, where the percentage of the number of experts which simultaneously take the correct decision is reported. In particular, Figure 7a details the percentage of the number of experts correctly assigning the class fire: we can note that all the experts correctly recognize the fire in most of the situations (69%), while two experts assign the class fire in the remaining 31%.

The advantage in using a MES is much more evident in Figure 7b, which refers to non fire videos. In this case, only 17% of videos are correctly classified by all the experts. On the other hand, most of the videos (61%) are assigned to the correct class by two experts, so confirming the successful combination obtained thanks to the proposed approach.

In order to better appreciate the behavior described above,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:10.1109/TCSVT.2015.2392531, IEEE Transactions on Circuits and Systems for Video Technology

TABLE I: The dataset used for the experimentation.

Video	Resolution	Frame Rate	Frames	Fire	Notes
Fire1	320x240	15	705	yes	A fire generated into a bucket and a person walking near it. Video downloaded from [34].
Fire2	320x240	29	116	yes	A fire very far from the camera generated into a bucket. The video has been downloaded from [34].
Fire3	400x256	15	255	yes	A big fire in a forest. The video has been acquired by [35] and downloaded from [34].
Fire4	400x256	15	240	yes	See the notes of the video *Fire3*.
Fire5	400x256	15	195	yes	See the notes of the video *Fire3*.
Fire6	320x240	10	1200	yes	A fire generated in a red ground. Video downloaded from [34].
Fire7	400x256	15	195	yes	See the notes of the video *Fire3*.
Fire8	400x256	15	240	yes	See the notes of the video *Fire3*.
Fire9	400x256	15	240	yes	See the notes of the video *Fire3*.
Fire10	400x256	15	210	yes	See the notes of the video *Fire3*.
Fire11	400x256	15	210	yes	See the notes of the video *Fire3*.
Fire12	400x256	15	210	yes	See the notes of the video *Fire3*.
Fire13	320x240	25	1650	yes	A fire in a bucket in indoor environment. Video downloaded from [34].
Fire14	320x240	15	5535	yes	Fire generated by a paper box. The video has been acquired by the authors near a street.
Fire15	320x240	15	240	no	Some smoke seen from a closed window. A red reflection of the sun appears on the glass. Video downloaded from [34].
Fire16	320x240	10	900	no	Some smoke pot near a red dust bin. Video downloaded from [34].
Fire17	320x240	25	1725	no	Some smoke on the ground near a moving vehicle and moving trees. Video downloaded from [34].
Fire18	352x288	10	600	no	Some far smoke on a hill. Video downloaded from [34].
Fire19	320x240	10	630	no	Some smoke on a red ground. Video downloaded from [34].
Fire20	320x240	9	5958	no	Some smoke on a hill with red buildings. Video downloaded from [34].
Fire21	720x480	10	80	no	Some smoke far from the camera behind some moving trees. Video downloaded from [34].
Fire22	480x272	25	22500	no	Some smoke behind a mountain in front of the University of Salerno. The video has been acquired by the authors.
Fire23	720x576	7	6097	no	Some smoke above a mountain. The video has been downloaded from [34].
Fire24	320x240	10	342	no	Some smoke in a room. Video downloaded from [34].
Fire25	352x288	10	140	no	Some smoke far from the camera in a city. Video downloaded from [34].
Fire26	720x576	7	847	no	See the notes of the video *Fire24*.
Fire27	320x240	10	1400	no	See the notes of the video *Fire19*.
Fire28	352x288	25	6025	no	See the notes of the video *Fire18*.
Fire29	720x576	10	600	no	Some smoke in a city covering red buildings. Video downloaded from [34].
Fire30	800x600	15	1920	no	A person moving in a lab holding a red ball. The video has been acquired by the authors.
Fire31	800x600	15	1485	no	A person moving in a labwith a red notebook. The video has been acquired by the authors.

a few examples are shown in Figure 6; in Figure 6a the fire is correctly recognized by all the experts: the color respects all the rules, the shape variation in consecutive frames is consistent and the movement of the corner points detected is very disordered. A different situation happens in Figure 6b, where the only classifier detecting the fire is the one based on the color: in this case, the uniform movement of the salient points associated to the ball as well as its constant shape allow the MES to avoid a false positive introduced by the use of a single expert. In Figures 6c and 6d other two examples are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:10.1109/TCSVT.2015.2392531, IEEE Transactions on Circuits and Systems for Video Technology

(a) fire1 (b) fire2

(c) fire4 (d) fire6

(e) fire13 (f) fire14

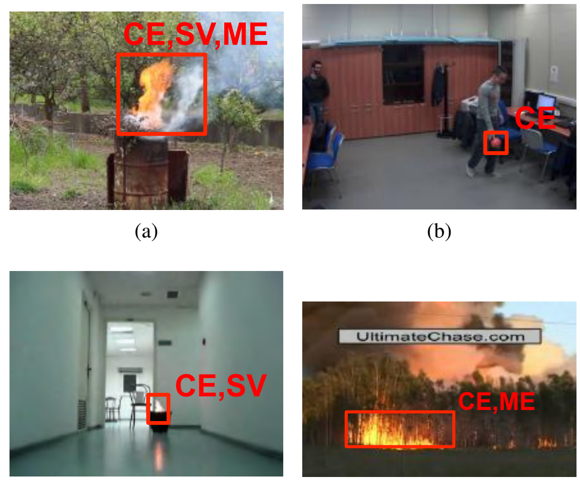
(g) fire15 (h) fire16

(i) fire19 (j) fire20

(k) fire22 (l) fire31

Fig. 5: Examples of images extracted from the videos used for testing the method.

shown: in particular, in the former a small fire with a variable shape has both a uniform color and a uniform movement of the salient points. The combination of color and shape variation



(a) (b)

(c) (d)

Fig. 6: The three experts in action; the red box indicates the position of the fire, while the letter on it refers to the expert recognizing the presence of the fire.

experts helps the proposed system to correctly detect the fire. The last example shows a very big but settled fire, whose shape is stable and so it is not useful for the detection. In this situation, the combination of the experts based on color and motion allows the MES to take the correct decision about the presence of fire.

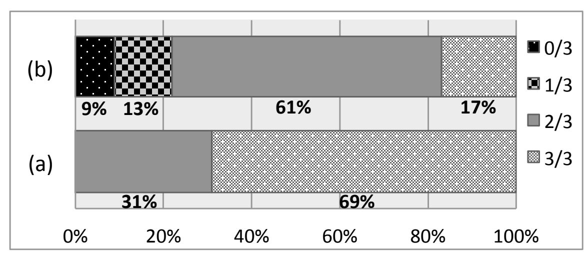


Fig. 7: Number of experts simultaneously taking the correct decision in fire (a) and non fire (b) videos. For instance, 31% of situations are correctly assigned to the class fire by two experts over three while in the remaining 69% all the three experts correctly recognize the fire.

Table II also shows a comparison of the proposed approach with three recent, state-of-the art methodologies ([15], [16], [29]) that have been chosen because they too are based on the combined use of color, motion and shape information. For [15] we have used two different versions: the original one, which, as proposed by its authors, analyzes the images in the RGB color space, and a version modified by us, working instead in YUV space; we have chosen this modification on the base of [7], where it is shown experimentally that color-based methods work better in YUV than in RGB, as confirmed by the results in Table II. The table shows that methods based on the combination of different kinds of information significantly outperform the single experts in terms of False Positives; the difference in terms of False Negatives is not so strong. Thus

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:10.1109/TCSVT.2015.2392531, IEEE Transactions on Circuits and Systems for Video Technology

TABLE II: Comparison of the proposed approach with state of the art methodologies in terms of Accuracy, False Positives and False Negatives.

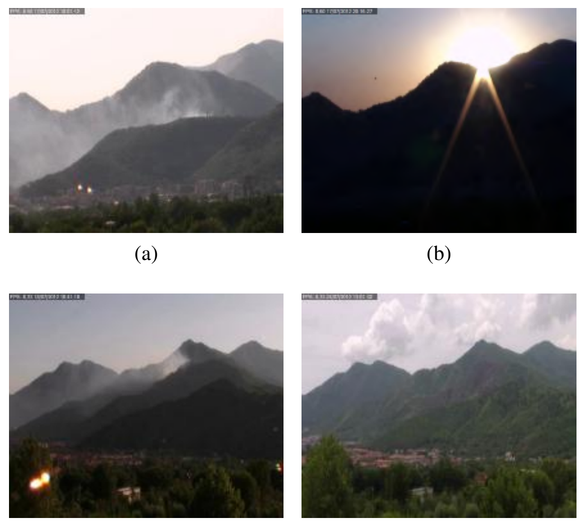
Typology	Method		Accuracy	False Positives	False Negatives
Single Expert	**CE**	[7]	83.87 %	29.41 %	0 %
	ME	Proposed	71.43 %	53.33 %	0 %
	SV		53.57 %	66.67 %	21.85 %
MES	**CE + SV**		88.29 %	13.33 %	9.74 %
	CE + ME	[25]	92.86 %	13.33 %	0 %
	CE + ME + SV	Proposed	**93.55 %**	11.76 %	0 %
Other Methods	**RGB+Shape+Motion**	[15]	74.20 %	41.18 %	7.14 %
	YUV+Shape+Motion	[15]	87.10 %	17.65 %	7.14 %
	Color+Shape+Motion	[16]	90.32 %	**5.88 %**	14.29 %
	Color+Shape+Motion	[29]	87.10 %	11.76 %	14.29 %

the combination helps more to improve the specificity than the sensitivity of the system. The proposed approach overcomes all the other considered methodologies [15] [16] [29] in terms of accuracy (93.55% against 89.29%, 90.32% and 87.10%, respectively). On the other hand, the best method in terms of False Positives is [16] (11.76% of the proposed approach with respect to 5.88% of [16]). The better False Positive Rate of [16] is however balanced by an improved False Negative Rate of our method, which shows no False Negatives (i.e. no fires are missed) versus a 14.29% False Negative Rate of [16]. While the difference between the two algorithms in terms of accuracy may seem not very large, the differences in the distribution of False Positives and False Negatives can make each of the two methods preferable depending on the requirements of the specific application.

A more detailed comparison for each of the considered videos is shown in Table I of the Electronic Annex: we can note that, differently from the other considered approaches, our method achieves a 100% true positive rate, since it is able to also retrieve very small flames (as the ones in videos fire1, fire2, fire6 or fire13). This is mainly due to the introduction of the MES for taking the final decision about the event, which is able to detect the onset of small fires at an early stage, when the amount of motion is still not very large. It is also evident that the method [16] is impressive for its reduced false positive rate, causing on the whole dataset just a single False Positive.

In order to further confirm the effectiveness of the proposed approach, we also evaluated it over a second freely available dataset (hereinafter D2)[2]. It is composed by 149 videos, each lasting approximatively 15 minutes, so resulting in more than 35 hours of recording; D2 contains very challenging situations, often recovered as fire by traditional color based approaches: red houses in a wide valley (see Figures 8a and 8d), a mountain at sunset (see Figure 8b) and lens flares (bright spots due to reflections of the sunlight on lens surfaces, see Figures 8a and 8c).

Although the situations are very challenging, no false positives are detected by our MES. The result is very encouraging, especially if compared with CE, achieving on the same dataset 12% of false positives. It is worth pointing out that such errors are localized in approximatively 7 hours, mainly at sunset, and are due to lens flares. Such typology of errors is completely solved by the proposed approach, able to take advantage of the disordered movement of the flames.



(a) (b)

(c) (d)

Fig. 8: Some examples of the Dataset D2, showing red houses in the wide valley, the mountain at sunset and some lens flares.

Finally, we have also evaluated the computational cost of the proposed approach over two very different platforms: the former is a traditional low-cost computer, equipped with an Intel dual core T7300 processor and with a RAM of 4GB. The latter is a Raspberry Pi B, a Broadcom BCM2835 System-on-a-chip (SoC), equipped with an ARM processor running at 700 MHz and with a RAM of 512 Mb. The main advantage in using such device lies in its affordable cost, around 35 dollars.

The proposed method is able to work, considering 1CIF videos, with an average frame rate of 60 fps and 3 fps respectively over the above mentioned platforms. Note that 60 fps is significantly higher than the traditional 25 - 30 fps that a

[2]The whole dataset can be downloaded from our website: *http://mivia.unisa.it/datasets/video-analysis-datasets/smoke-detection-dataset/*.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:10.1109/TCSVT.2015.2392531, IEEE Transactions on Circuits and Systems for Video Technology

traditional camera can reach during the acquisition. It implies that the proposed approach can be easily and very effectively used on existing intelligent video surveillance systems without requiring additional costs for the hardware needed for the images processing.

In order to better characterize the performance of the proposed approach, we also evaluated the time required by the different modules, namely the three experts (CE, ME and SV) and the module in charge of updating the background, extracting the foreground mask and labeling the connected components (FM). The contribution of each module is highlighted in Figure 9: the average time required to process the single frame has been computed and the percentage of each module with respect to the total time is reported. We can note that SV only marginally impacts on the execution time; this is due to the fact that the search of the minimum bounding boxes enclosing the blobs and of its properties (in terms of perimeter and area) is a very low-cost operation. Although the introduction of SV only slightly increases the performance of the MES (from 92.86% to 93.55% in terms of accuracy), the small additional effort strongly justifies its introduction in the proposed MES.

On the other side, the higher impacts are due to ME and CE: as for the former (85%), it is evident that the computation of the salient points, as well as their matching, is a very onerous operation. As for the latter, it may appear surprising the big effort required by the CE with respect to FM (CE: 11%, FM: 2%). It is worth pointing out that FM's operations (such as background updating and connected component labeling) are very common in computer vision, and thus very optimized versions have been proposed in standard libraries such as OpenCV.

Finally, it is worth pointing out that the computation time is strongly dependent on the particular image the algorithm is processing. In fact, it is evident that pixel-based modules (such as FM and CE) need to process the whole image independently of the objects moving inside. On the other hand, it is evident that the more are the objects moving inside the scene, the higher is the effort required by FM for detecting and analyzing the salient points. It implies that the variance with respect to the overall time required for the computation is about 51% of the overall time. Note that the final combination of the decisions taken by the three experts has not been considered, since the time required is very small with respect to the other modules.

In conclusion, the obtained results, both from a quantitative and a computational point of views, are very encouraging since they allow the proposed approach to be profitably used in real environments.

V. Conclusions

In this paper we propose a fire detection system using an ensemble of experts based on information about color, shape and flame movements. The approach has been tested on a wide database with the aim of assessing its performance both in terms of sensitivity and specificity. Experimentation confirmed the effectiveness of the MES approach, which allows to

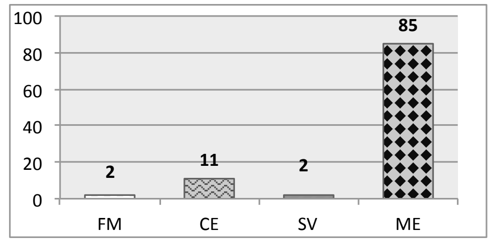


Fig. 9: The average execution time of our algorithm, in terms of percentage of the total time for any expert (CE,SV e ME) and the preliminary low level vision elaborations (FM).

achieve better performance in terms of true positive rate with respect to any of its composing experts. Particularly significant by the applicative point of view is its drastic reduction of false positives, from 29.41% to 11.76%. The comparison with other state of the art methods highlights its highest accuracy on the considered dataset; at the same time [16] although having a less true positive further reduces the false positive rate, conquering the best rank on this aspect. These considerations seem to suggest us that further investigations should be done in the direction of understanding how to put together the strength of the high accuracy of our method together with the strength of [16] of a really low false positive rate.

As for the execution efficiency, even though the system is made of three experts working simultaneously, its overall computational load is compatible with low cost embedded systems such as the Raspberry Pi; a carefully tuned implementation runs in real time at a frame rate of about 3 fps on images at a resolution of 320x240. A noteworthy additional outcome of this work is the database prepared for experimentations; a wide collection of videos containing fires filmed in different conditions and environments is completed with a significant number of non-fire videos, carefully selected among those highly confusable with scenes of fire. This made it possible to check the robustness of the system with respect to the generation of false positives. The database is publicly available at: `http://mivia.unisa.it/datasets/` .

Future work will be devoted to the integration in the same MES framework of a smoke detection algorithm, and to the extension of the approach to operating conditions currently not covered, such as its execution directly on board of the camera, and its use on Pan-Tilt-Zoom cameras.

Acknowledgments

This research has been partially supported by A.I.Tech s.r.l. (http://www.aitech-solutions.eu).

References

[1] A. E. Cetin, K. Dimitropoulos, B. Gouverneur, N. Grammalidis, O. Gunay, Y. H. Habiboglu, B. U. Toreyin, and S. Verstockt, "Video fire detection: a review," *Digital Signal Processing*, vol. 23, no. 6, pp. 1827 – 1843, 2013.

[2] Z. Xiong, R. Caballero, H. Wang, A. Finn, and P.-y. Peng, "Video fire detection: Techniques and applications in the fire industry," in *Multimedia Content Analysis*, ser. Signals and Communication Technology, A. Divakaran, Ed. Springer US, 2009, pp. 1–13.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:10.1109/TCSVT.2015.2392531, IEEE Transactions on Circuits and Systems for Video Technology

[3] T. Celik, H. Demirel, H. Ozkaramanli, and M. Uyguroglu, “Fire detection using statistical color model in video sequences,” *J. Vis. Comun. Image Represent.*, vol. 18, no. 2, pp. 176–185, Apr. 2007.

[4] H.-Y. J. Yoon-Ho Kim, Alla Kim, “Rgb color model based the fire detection algorithm in video sequences on wireless sensor network,” *International Journal of Distributed Sensor Networks*, 2014.

[5] C. Yu, Z. Mei, and X. Zhang, “A real-time video fire flame and smoke detection algorithm,” *Procedia Engineering*, vol. 62, no. 0, pp. 891 – 898, 2013, 9th Asia-Oceania Symposium on Fire Science and Technology.

[6] X. Qi and J. Ebert, “A computer vision-based method for fire detection in color videos,” *International Journal of Imaging*, vol. 2, no. 9 S, pp. 22–34, 2009.

[7] T. Celik and H. Demirel, “Fire detection in video sequences using a generic color model,” *Fire Safety Journal*, vol. 44, no. 2, pp. 147–158, 2009.

[8] T. Celik, H. Ozkaramanli, and H. Demirel, “Fire pixel classification using fuzzy logic and statistical color model,” in *ICASSP*, vol. 1, April 2007, pp. I–1205–I–1208.

[9] B. C. Ko, K.-H. Cheong, and J.-Y. Nam, “Fire detection based on vision sensor and support vector machines,” *Fire Safety Journal*, vol. 44, no. 3, pp. 322 – 329, 2009.

[10] M. Mueller, P. Karasev, I. Kolesov, and A. Tannenbaum, “Optical flow estimation for flame detection in videos,” *IEEE Trans. Image Process*, vol. 22, no. 7, pp. 2786–2797, July 2013.

[11] A. Rahman and M. Murshed, “Detection of multiple dynamic textures using feature space mapping,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 5, pp. 766–771, May 2009.

[12] B. C. Ko, S. J. Ham, and J.-Y. Nam, “Modeling and formalization of fuzzy finite automata for detection of irregular fire flames,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 12, pp. 1903–1912, Dec 2011.

[13] B. U. Töreyin, Y. Dedeoğlu, U. Güdükbay, and A. E. Çetin, “Computer vision based method for real-time fire and flame detection,” *Pattern Recogn. Lett.*, vol. 27, no. 1, pp. 49–58, Jan. 2006.

[14] P. Borges and E. Izquierdo, “A probabilistic approach for vision-based fire detection in videos,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 5, pp. 721–731, May 2010.

[15] A. Rafiee, R. Tavakoli, R. Dianat, S. Abbaspour, and M. Jamshidi, “Fire and smoke detection using wavelet analysis and disorder characteristics,” in *ICCRD*, vol. 3, March 2011, pp. 262–265.

[16] Y. Habiboglu, O. Gunay, and A. Cetin, “Covariance matrix-based fire and flame detection method in video,” *Mach. Vision Appl.*, vol. 23, no. 6, pp. 1103–1113, November 2012.

[17] O. Gunay, K. Tasdemir, B. Ugur Toreyin, and A. Cetin, “Fire detection in video using lms based active learning,” *Fire Technology*, vol. 46, no. 3, pp. 551–577, 2010.

[18] V. N. Vapnik, “An overview of statistical learning theory,” *Trans. Neur. Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999.

[19] L. Lam and C. Y. Suen, “Optimal combinations of pattern classifiers,” *Pattern Recogn. Letters*, vol. 16, no. 9, pp. 945 – 954, 1995.

[20] J. Kittler, “Combining classifiers: A theoretical framework,” *Pattern Analysis and Applications*, vol. 1, no. 1, pp. 18–27, 1998.

[21] P. Soda, G. Iannello, and M. Vento, “A multiple expert system for classifying fluorescent intensity in antinuclear autoantibodies analysis,” *Pattern Anal. Appl.*, vol. 12, no. 3, pp. 215–226, Sep. 2009.

[22] M. D. Santo, M. Molinara, F. Tortorella, and M. Vento, “Automatic classification of clustered microcalcifications by a multiple expert system,” *Pattern Recogn*, vol. 36, no. 7, pp. 1467 – 1477, 2003.

[23] L.-L. Huang and A. Shimizu, “A multi-expert approach for robust face detection,” *Pattern Recogn*, vol. 39, no. 9, pp. 1695 – 1703, 2006.

[24] L. P. Cordella, M. D. Santo, G. Percannella, C. Sansone, and M. Vento, “A multi-expert system for movie segmentation,” in *MCS*. London, UK, UK: Springer-Verlag, 2002, pp. 304–313.

[25] A. Greco, R. D. Lascio, A. Saggese, and M. Vento, “Improving fire detection reliability by a combination of videoanalytics,” in *ICIAR*, 2014.

[26] D. Conte, P. Foggia, M. Petretta, F. Tufano, and M. Vento, “Meeting the application requirements of intelligent video surveillance systems in moving object detection,” in *Pattern Recognition and Image Analysis*, ser. LNCS. Springer Berlin Heidelberg, 2005, vol. 3687, pp. 653–662.

[27] L. He, Y. Chao, and K. Suzuki, “A run-based two-scan labeling algorithm,” *IEEE Trans. Image Process*, vol. 17, no. 5, pp. 749–756, May 2008.

[28] S. Tulyakov, S. Jaeger, V. Govindaraju, and D. S. Doermann, “Review of classifier combination methods.” in *Machine Learning in Document Analysis and Recognition*, ser. Studies in Computational Intelligence. Springer, 2008, vol. 90, pp. 361–386.

[29] T.-H. Chen, P.-H. Wu, and Y.-C. Chiou, “An early fire-detection method based on image processing,” in *ICIP*, vol. 3, Oct 2004, pp. 1707–1710 Vol. 3.

[30] C. Poynton, *Digital Video and HDTV Algorithms and Interfaces*, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003.

[31] T. Joachims, “Text categorization with suport vector machines: Learning with many relevant features,” in *ECML*. Springer-Verlag, 1998, pp. 137–142.

[32] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[33] E. Rosten and T. Drummond, “Machine learning for high-speed corner detection,” in *ECCV*, 2006, pp. 430–443.

[34] E. Cetin. (2014, July) Computer vision based fire detection dataset. [Online]. Available: http://signal.ee.bilkent.edu.tr/VisiFire/

[35] (2014, July) Ultimate chase. [Online]. Available: http://ultimatechase.com/

Pasquale Foggia received the Laurea degree (cum laude) in computer engineering in 1995, and the Ph.D. degree in electronic and computer engineering in 1999, from the Federico II University of Naples, Naples, Italy. He was an Associate Professor of computer science with the Department of Computer Science and Systems, University of Naples, from 2004 to 2008, and has been with the University of Salerno, Fisciano, Italy, since 2008. His current research interests include basic methodologies and applications in the fields of computer vision and pattern recognition. He is the author of several research papers on these subjects. Dr. Foggia has been a member of the International Association for Pattern Recognition (IAPR), and has been involved in the activities of the IAPR Technical Committee 15 (Graph-based Representations in Pattern Recognition) since 1997.

Alessia Saggese received in December 2010 the Laurea degree (cum laude) in Computer Engineering from the University of Salerno, Italy and in February 2014 the Ph.D. degree in electronic and computer engineering from the University of Salerno, Italy, and from the University of Caen Basse Normandie, France, with a thesis entitled Detecting and indexing moving objects for Behavior Analysis by Video and Audio Interpretation. She is currently a Research Fellow of the University of Salerno. Since July 2012 she is a member of the IAPR Technical Committee 15 (Graph-based Representations in Pattern Recognition). Her research interests mainly pertain real time video analysis and interpretation for traffic monitoring and video surveillance applications.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:10.1109/TCSVT.2015.2392531, IEEE Transactions on Circuits and Systems for Video Technology

Mario Vento is a fellow scientist of the International Association Pattern Recognition (IAPR). Currently he is Full Professor of Computer Science and Artificial Intelligence at the University of Salerno (Italy), where he is the coordinator of the Artificial Vision Lab. From 2002 to 2006 he served as chairman of IAPR Technical Committee TC15 on Graph Based Representation in Pattern Recognition, and from 2003 as associate editor of the Electronic Letters on Computer Vision and Image Analysis. His research interests fall in the areas of Artificial Intelligence, Image Analysis, Pattern Recognition, Machine Learning and Computer Vision. More specifically, his research activity covered real time video analysis and interpretation for traffic monitoring and video surveillance applications, Classification Techniques, either Statistical, Syntactic and Structural, Exact and Inexact Graph Matching, Multi-Expert Classification and Learning Methodologies for Structural Descriptions. He authored over 200 research papers in International Journals and Conference Proceedings and serves as referee for many relevant journals in the field of Pattern Recognition and Machine Intelligence.