Video summarization guiding evaluative rectification for industrial activity recognition

Athanasios S. Voulodimos
School of Electrical & Computer Engineering
National Technical University of Athens
thanosv@mail.ntua.gr

Anastasios D. Doulamis
Dept of Production Engineering & Management
Technical University of Crete
adoulam@dpem.tuc.gr

Dimitrios I. Kosmopoulos
Dept of Computer Science & Engineering
University of Texas at Arlington
kosmopo@uta.edu

Theodora A. Varvarigou
School of Electrical & Computer Engineering
National Technical University of Athens
dora@telecom.ntua.gr

Abstract

In this paper we present a video summarization method that extracts key-frames from industrial surveillance videos, thus dramatically reducing the number of frames without significant loss of semantic content. We propose to use the produced summaries as training set for neural network based Evaluative Rectification. Evaluative Rectification is a method that exploits an expert user's feedback regarding the correctness of an activity recognition framework on part of the data in order to enhance future classification results. The size of the training sample set usually depends on the topology of the network and on the complexity of the environment and activities observed. However, as is shown by the experiments conducted in a real-world industrial activity recognition dataset, using a much smaller but representative sample stemming from our summarization technique leads to significantly higher accuracy rates than those attained by a same size but randomly chosen set. To obtain comparable improvement in accuracy without the summarization technique, the experiments show that a far larger training sample set is needed, therefore requiring significantly increased human resources and computational cost.

1. Introduction

Event and activity recognition in video remains one of the most intriguing problems in the computer vision community. This is mainly due to the useful applications associated with automatic event analysis and detection, as well as to the interesting nature of the issue from a researcher's point of view. The difficulty of the problem directly depends on the challenges posed by the characteristics of the monitored environment and the complexity of the activities observed. Complicated environments such as the production line of industrial plants provide for a rather challenging basis for algorithms attempting action/activity/behavior recognition, since they often involve cluttered background, frequent illumination changes, severe occlusions, target deformations, outliers, etc. In these cases, detecting and tracking the salient scene objects [7], whether human, machinery or other, can become exceptionally hard. To this end, approaches that bypass these intermediate levels of semantic complexity through the use of local or holistic image-based features for scene representation have been proposed. For example, in [10] the authors propose a framework for robust visual behavior recognition in an industrial environment based on hidden Markov models (HMM) classifiers, multicamera fusion and the use of the multivariate Student's t-distribution as observation model for higher tolerance to outliers.

In [9] the concept of *Evaluative Rectification* is introduced in the context of automatic activity recognition. The proposed method is based on relevance feedback and tries to readjust the estimated probabilities generated by the HMM classifier by means of a maximum likelihood framework, in the direction of minimizing the overall misclassification error. This readjustment is accomplished by exploiting an expert user's interaction who evaluates the results of the task recognition process through a selection set of workflow tasks (i.e. targeted activities) that have been either correctly or erroneously classified by the HMMs approach along with the respective target class that these visually detected tasks belong to. It should be noted that in [9] the proposed method is applied to the likelihoods generated by an

HMM based framework; however, the readjustment can be performed regardless of the classifier employed. A learning strategy is used for dynamically modifying the probabilities of the classifier. In particular, a non-linear least square method is employed by the use of feedforward neural network structure. A feedforward neural network is able to approximate any continuous non-linear function with a certain degree of accuracy. The Evaluative Rectification method exploits this property of neural networks to estimate the non-linear modification of the probabilities extracted by the classifier, in order to fit the target ideal probabilities.

An important issue involved in the Evalutive Rectification method is the availability of training samples. The number of samples required for training depends on the topology of the neural network employed, which in turn depends on the number of the activities that are being modeled, as well as the high or low inter- and intraclass variance among them. A more complex environment with numerous challenging activities to be recognized is bound to call for a larger network with more neurons (or even additional hidden layers), thus requiring more training samples. This creates the need for more feedback given by the expert user, who should therefore be increasingly occupied with the feedback task. Nevertheless, the amount of available training data is not the sole factor that ought to be taken into consideration. What is equally or even more important is that the sample consisting of these training data, on which the feedback is given should be *representative*. In other words, if the expert user is called to provide feedback at specific time intervals (even relatively lengthy ones) where, e.g., the observed activity (or visual pattern) is always the same (or similar), then that particular activity would be overrepresented in the sample, and other activities would be underrepresented. This could lead to less successful training of the neural network, even though the absolute number of training samples could have been sufficiently large.

Our goal is therefore to propose a method to create a representative sample of frames, on which the user will be required to provide relevance feedback. Achieving this goal will induce better training of the neural network (and, hence, better overall recognition rates) with the size of training sample, and consequently the involvement of the expert user, being kept to a minimum. To this end, we propose the employment of a video summarization technique that significantly reduces the number of frames while minimizing the loss of semantic content. The key-frames extracted by the employment of this technique are the training samples to which the expert user provides relevance feedback in the context of the Evaluative Rectification approach, with a view to enhancing the overall recognition rates.

The remainder of this paper is structured as follows: Section 2 provides a brief overview of related work in the fields of video summarization and relevance feedback. Section 3 details the proposed video summarization method, while Section 4 briefly reminds the neural network based Evaluative Rectification method. In Section 5 we experimentally verify the applicability of the proposed methods in the context of an activity recognition framework in an industrial environment, while Section 6 concludes the paper with the lessons learned.

2. Related work

Summarization The first approaches towards an automatic video summarization scheme had the goal of extracting key-frames at regular time instances or within a shot [5]. Such a selection, however, is far from being representative especially when someone should quickly overview complex industrial processes as in our case. Thus, from the end of the 90's video summarization has boosted the multimedia research society [19]. Furthermore, research has been concentrated on specific type of video content like sports [16, 2], instructional videos [4] or even facial emotion recognition [6]. Optimal algorithms are also proposed maximizing entropy and exploiting information theoretic measures [11], [14], and/or perceptual users' centric video summaries [12].

The main obstacle of the above mentioned algorithms is that they fail in identifying key-frames when period movements take place while their computational complexities are not affordable especially for very long video sequences being continually captured, as the ones adopted in surveillance systems. For this reason, in our industrial application scenario we do not adopt any of the aforementioned techniques for extracting key-frames from the online captured industrial video monitoring. Instead, we propose another idea which exploits the fluctuation of the feature vector trajectory. This way, it is possible to detect periodic movements which are crucial for inspecting the products' assembly quality in manufactories. To identify the key points on the trajectory curve, we initially estimate the curvature of the feature trace via discrete second derivatives equaling zero. This approach yields fast implementation of the proposed scheme which can be embedded in industrial devices and thus allow a real-time implementation of the algorithm in the factory. Real-time summarization algorithms have been also proposed in [5] and [1] but in the presented technique we also ensure detection of periodic effects which are important for a quick but meaningful overview of industrial processes. To eliminate possible noise, the discrete second derivatives are smoothed over a time window.

Evaluative Rectification As has already been mentioned, evaluative rectification is inspired from relevance feedback. Relevance feedback is a common approach for automatically adjusting the response of a system regarding information taken from user's interaction [8]. Originally, it has been developed in traditional information retrieval systems [15],

but it has been now extended to other applications, such as surveillance systems [13], [3]. Relevance feedback is actually an online learning strategy which re-weights important parameters of a procedure in order to improve its performance. Re-weighting strategies can be linear or non-linear relying either on heuristic or optimized methodologies [8]. Linear and heuristic approaches usually adjust the degree of importance of several parameters involved in the selection process. Instead, non-linear methods adjust the applied method itself using function approximation strategies. In this direction neural network models have been introduced as non-linear function approximation systems. However, such approaches have been applied for information retrieval systems instead of surveillance applications. It is clear that it is not straightforward to move from one type of application to another due to the quite different requirements of both applications. A comprehensive review regarding algorithms of relevance feedback in image retrieval has been provided in [20]. In this paper, the authors compare different techniques of relevance feedback with respect to the type of training data, the adopted organization strategies, the similarity metrics used, the implemented learning strategies and finally the effect of negative samples in the training performance. Finally, in [9] the authors employ the relevance feedback mechanism in an endeavor to exploit expert user's feedback so as to improve the classification rates of an HMM based industrial activity recognition framework, thus proposing the Evaluative Rectification approach. However, as is explained in the Introduction, the issues of numerical sufficiency as well as representativeness arise, since in a realistic practical implementation, the expert user would provide feedback either at random or at specific time intervals. Our aim in this paper is therefore to "move" the selection of frames on which the feedback is to be given from the "time domain" to the "content domain", thus decreasing the number of frames required for successful training while attaining similar or better performance.

3. Summarization

Hereby we describe the approach adopted to summarize industrial videos being captured as a result of a visual surveillance system. The goal is to dramatically reduce visual information, (e.g., the number of frames), without, however, losing important information as far as the meaning of an industrial activity (called "task") and/or workflow is concerned. We claim that the semantic content is expressed in the sense of feature vector complexity. In this work we use holistic features based on Pixel Change History [18] and Zernike moments calculation; [10] contains a detailed description of the feature extraction process. Therefore other types of features may yield different summaries. However, one should assess the effectiveness and efficiency of a summarization algorithm, both in terms of performance and computational complexity, on the same feature elements for fairness.

The sum of the squared coefficients of the Zernike moments can express the motion energy or in other words a measure of motion in the current scene. It is defined as:

$$E = \sum_{p=0}^{Q} \sum_{\substack{q \le p \\ \frac{p-q}{2}:even}} \|A_{pq}\|^2 \tag{1}$$

where Q is the selected order of the moments and $\|\|$ the $L2$ norm. In other words, the energy of a frame is defined as the sum of the squared $L2$ norms of the associated Zernike moments up to the order Q. The total energy in a distributed camera setting can be defined as the weighted sum of the energies of the individual streams, which are given by (1).

The energy can be plotted for each video frame forming a trajectory, which expresses the temporal variation of the energy shape through time. Thus, selection of the most representative frames within a shot is equivalent to selection of appropriate curve points, able to represent the corresponding trajectory. In our case, the second derivative of the shape energy for all frames within a shot with respect to time is used as a curvature measure. Local maxima correspond to time instances of peak variation of the object shape. In addition, local minima indicate low variation of the object shape.

Let us also denote as $E(k)$ the energy of shape coefficients to the k-th frame of the examined shot. Initially, the first derivative of signal $E(k)$, say, is evaluated with respect to time index k. Since, however, variable k takes values in discrete time, the first derivative is approximated as the difference of shape energy between two successive frames.

However, the first derivative is sensitive to noise since differentiation of a signal stresses the high pass components. For this reason, a weighted average of the first derivative, say E'_w, over a time window, is used to make smoother the fluctuation for the magnitude of the frame feature vectors.

Particularly, the weighted first derivative is given as

$$E'_w(k) = \sum_{l=a_1(k)}^{l=\beta_1(k)} w_{l-k}(E(l+1) - E(l)), k = 0, ..., M-2 \tag{2}$$

where $\alpha_1(k) = max(0, k - N_w)$ and $\beta_1(k) = min(M - 2, k + N_w)$ and $2 * N_w + 1$ is the length of the window, centered at frame k. Variable M indicates the number of frames of the shot. It can be seen from (3) that the window length linearly reduces at shot limits.

The weights w_l are defined for $l \in \{-N_w, N_w\}$; in the simple case, all weights w_l are considered equal to each other, meaning that the derivatives of all frame feature vec-

tors within the window interval present the same importance,

$$w_l = \frac{1}{2N_w + 1}, l = -N_w, ..., N_w \quad (3)$$

Since frames are discrete time instances, we can model the derivative via difference equations and thus we can estimate the second derivative E''_w , for the k-th frame as:

$$E''_w(k) = \sum_{l=\alpha(k)}^{l=\beta(k)} w_{l-k} E''(l), k = 0, ..., M-3 \quad (4)$$

where $\alpha(k) = max(0, k - N_w)$ and $\beta(k) = min(M - 2, k+N_w)$, and $E''(k) = E'(k+1) - E'(k)$. The local maxima and minima of E'' are considered as appropriate curve points, i.e., as time instances for the selected key-frames.

Note that this algorithm is extremely fast since each process is implemented independently from frame to frame and the time required for applying the frame difference is minimal. Also the Zernike feature vectors are computed anyway for task rcognition purposes, so the total overhead is actually minimal.

Hence, the local maxima and minima can be estimated as the union of two sets $X = X_M \cup X_m$; the X_M contains the time instances of frames corresponding to the local maxima of E'', while the X_m the time instances of local minima of E''. The sets X_M and X_m are estimated as follows:

$$\begin{aligned} X_M &= \{k : E''(k-1) < E''(k) \& E''(k) > E''(k+1)\} \\ X_m &= \{k : E''(k-1) > E''(k) \& E''(k) < E''(k+1)\} \end{aligned} \quad (5)$$

4. Evaluative Rectification

In this section, we provide a brief overview of the Evaluative Rectification method. This method aims at dynamically correcting erroneous classifications of an activity recognition framework, such as the one described in [10], which is based on HMM classifiers and possibly multicamera fusion. We hereby denote as $\mathbf{p}_i$ the observation probability vector (generated by the classifier) the elements of which express the probability of the corresponding frame to belong to one of the, say, M available activity classes and as $\mathbf{d}_i$ the target vector containing the ideal probabilities for the i^{th} sample, i.e., all its elements are zero apart from one which is equal to one. The creation of $\mathbf{d}_i$ is based on the feedback provided by the expert user. Assuming that the non-linear relationship between vectors $\mathbf{p}_i$ and $\mathbf{d}_i$ is indicated by a vector function $\underline{f}(\cdot)$, we introduce a feedforward neural network model able to accurately approximate the unknown vector function $\underline{f}(\cdot)$ with a certain degree of accuracy. Let $\mathbf{w}$ be the weight vector that includes all the parameters (weights) of the non-linear neural network, which is the core of the readjustment mechanism. To estimate the weights $\mathbf{w}$ we need to apply a training algorithm, which actually minimizes the mean square error among all selected from the expert user data and the respective output of the network when a particular set of weights is applied. That is,

$$\mathbf{w} = \arg \min_{forall \mathbf{w}} \epsilon = \arg \min_{forall \mathbf{w}} \sum_i (\underline{f}_{\mathbf{w}}(\mathbf{p}_i) - \mathbf{d}_i)^2 \quad (6)$$

Expressing the unknown non-linear vector function as a feedforward neural network, we are able to estimate vector $\mathbf{w}$ that minimizes (6) using the backpropagation training algorithm. It is clear that the samples of the training set should be greater than the number of neural network parameters, that is the dimension of the weight vector $\mathbf{w}$. In this paper the selection process of the frames which will compose the training set will be based on the summarization method described in Section 3.

5. Experiments and Results

The experimental validation of the proposed approach has been performed on a very challenging dataset [17] acquired from the production line of a major automobile manufacturer.[1] The activity recognition framework used is the one presented in [10]. The workspace configuration and the cameras' positioning is given in Fig. 1. The workflow on this assembly line included tasks of picking several parts from racks and placing them on a designated cell some meters away, where welding was performed. More specifically, the behaviors we were aiming to model in the examined application are briefly the following:

1. One worker picks part #1 from rack #1 and places it on the welding cell.
2. Two workers pick part #2a from rack #2 and place it on the welding cell.
3. Two workers pick part #2b from rack #3 and place it on the welding cell.
4. A worker picks up parts #3a and #3b from rack #4 and places them on the welding cell.
5. A worker picks up part #4 from rack #1 and places it on the welding cell.
6. Two workers pick up part #5 from rack #5 and place it on the welding cell.
7. Welding: two workers grab the welding tools and weld the parts together.

[1] The dataset, the activities' labeling as well as the features used are publicly available through http://www.scovis.eu/.

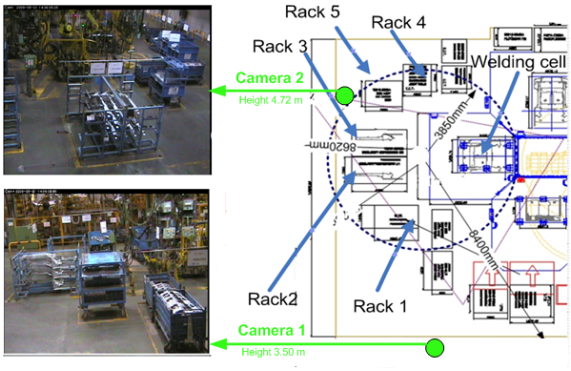


Figure 1. Depiction of workcell along with the position of the cameras and racks #1-5.

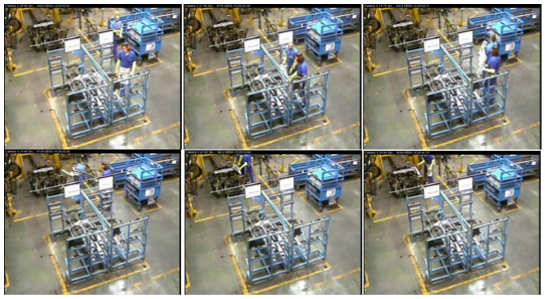


Figure 2. Typical execution of a task. Challenges of the industrial dataset include severe occlusions, cluttered background and frequent illumination changes among others.

Each of the above tasks/activities can be regarded as a class of behavioral patterns that has to be recognized. A sample task (task 2) is presented in Fig. 2. For our experiments, we have used 20 segmented sequences representing full assembly cycles, each one containing each of the seven behaviors/tasks. The total number of frames was approximately 80,000. The videos were shot by two PTZ cameras at an approximate framerate of 25 fps and at a resolution of 704×576. The annotation of these frames has been done manually.

5.1. Summarization

We evaluated the proposed video summarization algorithm in terms of effectiveness, i.e., how relevant the extracted key-frames were with respect to the overall meaning of an industrial task but also in terms of computational efficiency. For the latter issue, we stress that the adopted algorithm was implemented to run in real-time and during the frame acquisition. Thus, the proposed summarization approach imposed minimal computational burdens allowing its implementation in on-line video capturing devices. Furthermore, the algorithm can be implemented in industrial embedded devices fostering its applicability in industries.

The goal was to detect a small number of key-frames that were able to represent the whole content visual complexity occurred during the process, without requiring an industrial engineer to browse through the whole sequence in order to ensure whether a workflow is executed in accordance to safety and quality requirements. The experiments have been conducted for all data in the sequence. In Fig. 3 we present a typical example that concerns task 2, in scenario 1. The scene presents two workers collaborating in carrying a door of a vehicle. In this figure, we have depicted one per 20 frames so as to give a complete overview of the complexity of the scene. Fig. 3d shows the fluctuation of E over time. We observe that a high differentiated of the feature magnitude is noticed among frames 250 and 300 which presents the high activity of the workers. To eliminate feature noise in this plot, we low-passed the first signal deriving the smoothed feature magnitude as illustrated in Fig.3e. The zero-crossings of the 2nd derivative (after filtering) gave the key-frames. The number of key-frames was not predefined but was inherently estimated by the scheme with respect to the complexity of the visual content.

We have also objectively evaluated the proposed summarization algorithm over the whole dataset. For each task a small number of key-frames was extracted, as described, and shown to two industrial engineers. Then, they identified which the task was that has been executed by observing only this small number of key-frames. For all cases, the tasks have been correctly recognized by the expert users. This reveals that the proposed summarization scheme represented sufficiently the complexity and periodicity of the industrial visual content. Finally, the industrial engineers were asked to detect inappropriate executions that may yield to quality damages and/or security, safety alarm. More than 97% of the false processed were correctly detected meaning that the summarization algorithm extracts a very small but sufficiently representative set of key-frames that actually express the whole task complexity.

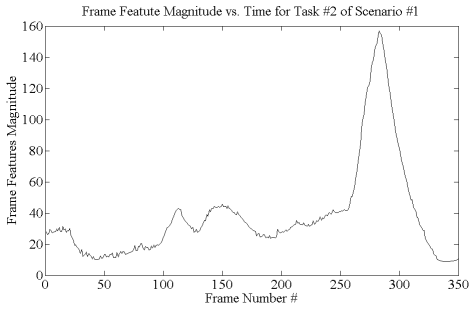
5.2. Evaluative Rectification

Here we experiment on the impact of the training sample set on the effectiveness of the Evaluative Rectification method. The samples are represented using the respective probability vector extracted by the HMM-based activity recognition framework, and the targeted correct classification of the frame, as provided by the expert user's feedback. For the HMM-based recognition half of the data were used for training and the other half for testing in each experimental setup. A feedforward neural network model is trained to adjust the probabilities extracted by the HMM in order to minimize the erroneous classifications. Regarding the structure of the feedfoward neural network, we use one hidden layer with fifteen (15) neurons. The neural network has seven input nodes and seven output nodes (equal to the number of modeled activities), thus making a total of 210 weights to be learned. The transfer function is the

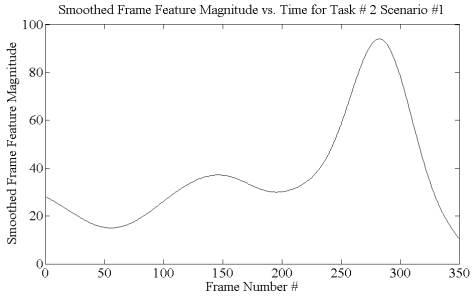
(a) frames 1-20-40-60-80-100

(b) frames 120-140-160-180-200-220

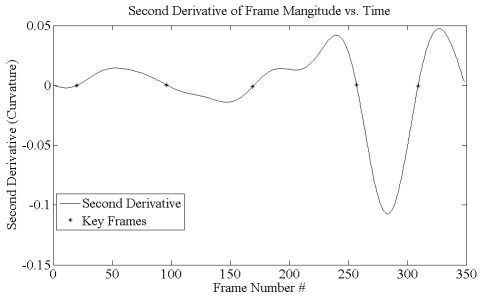
(c) frames 240-260-280-300-320-340



(d) First derivative of E



(e) Filtered first Derivative of E



(f) Zero crossings of 2nd Derivative

(g) Selected key frames: 20, 96, 169, 257, 309

Figure 3. Extraction of keyframes for scenario 1, task 1, camera 1

sigmoid. The experiments are conducted in the context of four different experimental setups (ES). Each experimental setup refers to a different part of the dataset and/or a different camera stream, hence the differences in the accuracy rates. Each set of experiments has been repeated ten times for reliability reasons and the presented results refer to the average accuracy rates and sample sizes across all runs.

Firstly, we compare the results of the employment of the Evaluative Rectification method with training sample sets that comprise approximately 5% of the total number of frames on average: in one case, the frames composing the training set are chosen at random; in the other case, the frames selected are the ones indicated as key-frames by the summarization technique described above. The results of these experiments in terms of overall average accuracy attained in the context of the four experimental setups are

shown in Table 1. We observe that in the case of a training set consisting of randomly chosen frames that correspond to 5% of the total number of frames on average, the attained accuracy rates are either lower than those achieved solely by the HMM-based framework (when the rates are already high) or slightly higher (when there is more room for improvement). These results are to be expected since the number of training samples is particularly small, thus hindering the successful training of the neural network. On the contrary, in the case where the summarization key-frames are used as training sample set, the accuracy rates are significantly higher than those attained by the mere use of HMM, although the size of the training set is roughly the same (and rather small), thus reflecting the importance of the representativeness in the training sample.

The limited number of training data in the context of the randomly chosen sample set is responsible for the unsatisfactory performance of the method in that case, which came as no surprise. It is interesting, nevertheless, to examine the performance of Evaluative Rectification with a randomly chosen sample set of a sufficiently large size, e.g. 35% of the total number of frames, and compare it to the one attained when using the much smaller summarization training set. It should be noted here that the stated sample sizes (35% or 5%) refer to the average sizes across all the experiment runs. The number of frames can in some cases fluctuate to slightly lower or higher values than the average ones. The related results are displayed in Table 2. As can be observed, the accuracy achieved with the use of the larger training set is slightly higher than the one yielded when using the far smaller summaries as training data. From this, we can infer that using the key-frames extracted through the summarization process described in Section 3 as training sample set leads to a significant enhancement of the HMM-based recognition rates which is comparable (only slightly inferior) to the improvement obtained when using a conventionally chosen but much larger training set. Therefore the expert user can now provide feedback in far fewer frames, thus considerably saving both human resources and computational cost, with a minimal sacrifice in overall performance.

Fig. 4 displays the average classification error for each of the approaches examined. Again, we can observe that the employment of evaluative rectification with a small training set consisting of summary key-frames introduces a significant improvement (in terms of error decrease this time) in comparison to the sole use of the HMM-based activity recognition framework. Furthermore, this improvement is comparable to the enhancement in the case of a far larger (by seven times on average) training set which has been selected ”conventionally”.

Table 1. Accuracy rates attained by the HMM-based ARF (i) solely, without ER, (ii) with ER where training sample set is selected randomly, (iii) with ER where training sample set is created by the summarization process (training sets are of equal size, i.e. approximately 5% of the total number of frames).

	ES 1	ES 2	ES 3	ES 4
HMM solely	81.1	84.3	50.2	51.8
HMM+ER (random set)	75.0	76.8	51.9	53.4
HMM+ER (summ. set)	86.2	88.9	58.7	60.2

Table 2. Accuracy rates attained in the case of (i) ER where training sample consists of summarization key-frames and (ii) ER where training sample set is selected randomly but is of significantly larger size.

HMM+ER	ES 1	ES 2	ES 3	ES 4
summaries small set	86.2	88.9	58.7	60.2
random large set	86.9	89.5	59.6	61.4

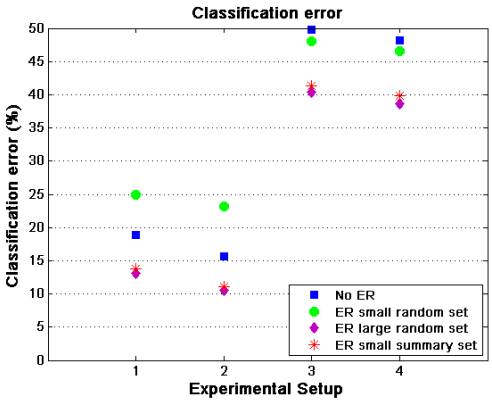


Figure 4. Classification error in four configurations (i) No Evaluative Rectification (ER) performed, (ii) ER with a small random training sample set, (iii) ER with a small training set consisting of summary key-frames and (iv) ER with a large random training sample set.

6. Conclusion

In this paper we have described a video summarization technique that extracts key-frames from surveillance videos with a view to using the provided summaries as training sample set in the context of the neural network based Evaluative Rectification method. Building upon an industrial activity recognition framework, the Evaluative Rectification method exploits feedback provided by an expert user in part of the data regarding the correctness of the classification performed. The complexity of the industrial environment and of the modeled activities influences the topology of the neural network employed and consequently the requirements on training data. As was shown, the number of training data is not the sole factor, since the representativeness of the chosen frames also plays a significant role. In particular, using the key-frames produced by our proposed

summarization technique as training set, leads, despite its small size, to an important enhancement in overall accuracy of the event/activity recognition framework, comparable to the improvement achieved when using a far larger training set (roughly seven times the size) chosen randomly. This ensures a significant saving of resources, since the expert user provides feedback not in the "time domain" any more, but in the more effective "content domain".

Acknowledgments. The research leading to these results has been supported by European Union funds and national funds from Greece and Cyprus under the project "POSEIDON: Development of an Intelligent System for Coast Monitoring using Camera Arrays and Sensor Networks" in the context of the inter-regional programme INTERREG (Greece-Cyprus cooperation) - contract agreement K1_3_10–17/6/2011.

References

[1] V. Anagnostopoulos, N. Doulamis, and A. Doulamis. Edge-motion video summarization: Economical video summarization for low powered devices. In *Image Analysis for Multimedia Interactive Services, 2009. WIAMIS '09. 10th Workshop on*, pages 284 –287, may 2009. 2

[2] F. Chen and C. De Vleeschouwer. Formulating team-sport video summarization as a resource allocation problem. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21(2):193 –205, feb. 2011. 2

[3] Z. Chengcui, C. Wei-Bang, C. Xin, Y. Lin, and J. John. A multiple instance learning and relevance feedback framework for retrieving abnormal incidents in surveillance videos. *Journal of Multimedia*, 5:310–321, 2010. 3

[4] C. Choudary and T. Liu. Summarization of visual content in instructional videos. *Multimedia, IEEE Transactions on*, 9(7):1443 –1455, nov. 2007. 2

[5] A. Doulamis, N. Doulamis, and S. Kollias. Non-sequential video content representation using temporal variation of feature vectors. *Consumer Electronics, IEEE Transactions on*, 46(3):758 –768, aug 2000. 2

[6] N. Doulamis. Context-adaptive and user-centric facial emotion classification. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 2, pages II – 53–6, sept. 2005. 2

[7] N. Doulamis. Coupled multi-object tracking and labeling for vehicle trajectory estimation and matching. *Multimedia Tools and Applications*, 50(1):173–198, 2010. 1

[8] N. Doulamis and A. Doulamis. Evaluation of relevance feedback schemes in content-based retrieval systems. *Signal Processing: Image Communication*, 21(4):334 – 357, 2006. 2, 3

[9] N. Doulamis, A. Voulodimos, D. Kosmopoulos, and T. Varvarigou. Enhanced human behavior recognition using hmm and evaluative rectification. In *ACM Multimedia, ARTEMIS Workshop*, 2010. 1, 3

[10] D. Kosmopoulos and S. Chatzis. Robust visual behavior recognition. *Signal Processing Magazine, IEEE*, 27(5):34 –45, sep. 2010. 1, 3, 4

[11] Z. Li, G. Schuster, and A. Katsaggelos. Minmax optimal video summarization. *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(10):1245 – 1256, oct. 2005. 2

[12] D. Liu, G. Hua, and T. Chen. A hierarchical visual model for video object summarization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(12):2178 –2190, dec. 2010. 2

[13] A. Oerlemans, J. T. Rijsdam, and M. S. Lew. Real-time object tracking with relevance feedback. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 101–104, New York, NY, USA, 2007. ACM. 3

[14] C. Panagiotakis, A. Doulamis, and G. Tziritas. Equivalent key frames selection based on iso-content principles. *Circuits and Systems for Video Technology, IEEE Transactions on*, 19(3):447 –451, march 2009. 2

[15] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The Smart retrieval system - experiments in automatic document processing*, pages 313–323. Englewood Cliffs, NJ: Prentice-Hall, 1971. 2

[16] D. Tjondronegoro, Y.-P. P. Chen, and B. Pham. Highlights for more complete sports video summarization. *Multimedia, IEEE*, 11(4):22 – 37, oct.-dec. 2004. 2

[17] A. Voulodimos, D. Kosmopoulos, G. Vasileiou, E. Sardis, A. Doulamis, V. Anagnostopoulos, C. Lalos, and T. Varvarigou. A dataset for workflow recognition in industrial scenes. In *Proc. of the 18th IEEE International Conference on Image Processing (ICIP) 2011*, pages 3310–3313, 2011. 4

[18] T. Xiang and S. Gong. Beyond tracking: modelling activity and understanding behaviour. *International Journal of Computer Vision*, 67:21–51, 2006. 3

[19] M. Yeung and B.-L. Yeo. Video visualization for compact presentation and fast browsing of pictorial content. *Circuits and Systems for Video Technology, IEEE Transactions on*, 7(5):771 –785, oct 1997. 2

[20] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8(6):536–544, 2003. 3