Multimodal Approach for Video Surveillance Indexing and Retrieval

Ali WALI, Adel M. ALIMI
REGIM: REsearch Group on Intelligent Machines, University of Sfax
National School of Engineers (ENIS), BP 1173, Sfax, 3038, Tunisia
ali.wali@ieee.org, adel.alimi@ieee.org

ABSTRACT: *In this paper, we present an overview of a new approach to indexing and searching the video sequence by the content that has been developed within the REGIMVid project. The platform termed MAVSIR provides High-level feature extraction from audio-visual content and concept/event-based video retrieval. We describe the architecture of the system as well as provide an overview of the descriptors supported to date. We then demonstrate the usefulness of the toolbox in the context of feature extraction, concepts/events learning and retrieval in large collections of video surveillance dataset.*

1. Introduction

Image and video indexing and retrieval continue to be an extremely active area within the broader multimedia research community [4], [13]. Interest is motivated by the very real requirement for efficient techniques for indexing large archives of audiovisual content in ways that facilitate subsequent usercentric accessing. Such a requirement is a by-product of the decreasing cost of storage and the now ubiquitous nature of capture devices. The result of which is that content repositories, either in the commercial domain (e.g. broadcasters or content providers repositories) or the personal archives are growing in number and size at virtually exponential rates. It is generally acknowledged that providing truly efficient usercentric access to large content archives requires indexing of the content in terms of the real world semantics of what it represents.

Furthermore, it is acknowledged that real progress in addressing this challenging task requires key advances in many complementary research areas such as; scalable coding of both audiovisual content and its metadata, database technology and user interface design. The REGIMVid project integrates many of these issues.In figure 1 we present our REGIMVid subsystem. A key effort within the project is to link audio-visual analysis with concept reasoning in order to extract semantic information. In this context, high-level preprocessing is necessary in order to extract descriptors that can be subsequently linked to the concept and used in the reasoning process. In addition to concept-based reasoning, the project has other research activities that require highlevel feature extraction (e.g. semantic summary of metadata [7], Text-based video retrieval [11], [8], event detection [1] and Semantic Access to Multimedia Data [6]) it was decided to develop a common platform for descriptor extraction that could be used throughout the project. In this paper, we describe Fig. 1. Our REGIMVid subsystem Architecture our subsystem for video surveillance indexing and retrieval. The remainder of the paper is organised as follows: a general overview of the toolbox is provided in Section 2, include a description of the architecture. In section 3 we present our approach to detect and extract of moving objects from video surveillance dataset. It includes a presentation of different concepts taken care by our system.We present the combining single SVM classifier for learning video event/concept in section 4. The descriptors of the visual feature extraction will be presented in section 5. Finally, we present our experimental results for both event and concept detection future plans for both the extension of the toolbox and its use in different scenarios.
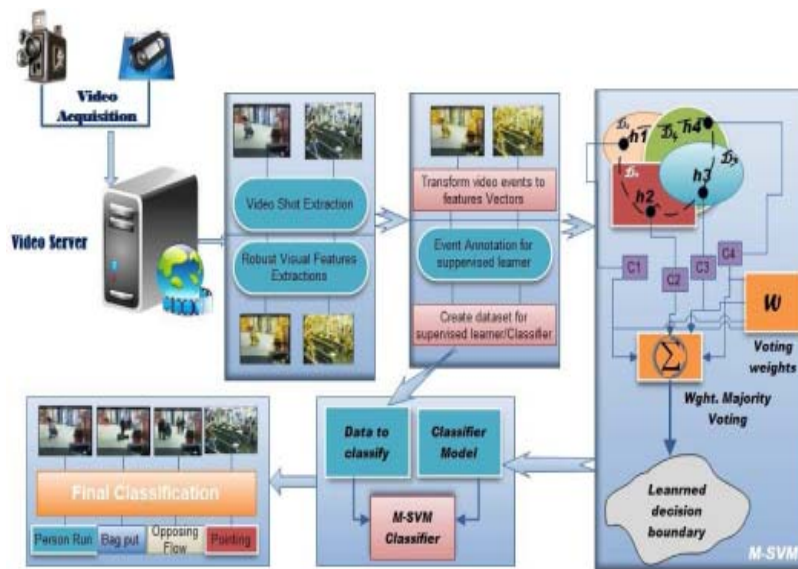
Figure 1. Our REGIMVid subsystem Architecture

2. Mavsir Toolbox For Video Surveillance Indexing Overview

In this section, we present an overview of the structure of the toolbox. The MAVSIR Toolbox currently supports extraction of 10 low-level (see section 5) visual descriptors. The design is based on the architecture of the MPEG-7 eXperimentation Model (XM), the official reference software of the ISO/IEC MPEG-7 standard.

The main objective of our system is to provide automatic content analysis using concept/event-based and low-level features. The system (figure 2) first detect and segment the moving object from video surveillance dataset. In the second step, it extracts three class of features from the each frame, from a static background and the segmented objects(the first class from Ω_{in}, the second from Ω_{out} and the last class is from each key-frame in RGB color space,see subsection 3.2), and labels them based on corresponding features. For example, if three features are used (color, texture and shape), each frame has at least three labels from Ω_{out}, three labels from Ω_{in} and three labels from key-frame.

This reduces the video as a sequence of labels containing the common features between consecutive frames. The sequence of labels aim to preserve the semantic content, while reducing the video into a simple form. It is apparent that the amount of data needed to encode the labels is an order of magnitude lower than the amount needed to encode the video itself. This simple form allows the machine learning techniques such as Support Vector Machines to extract high-level features.
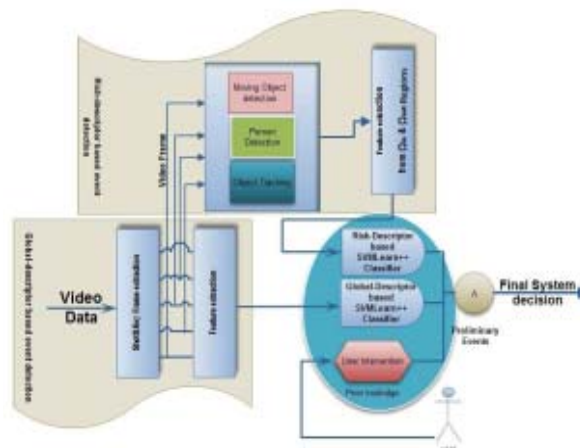

Figure 2. Overview of our system for video input

Our method offer a way to combine low-level features witch enhances the system performance. The high-level features extraction system according to our toolkit provides an open framework that allows easy integration of new features. In addition, the Toolbox can be integrated with traditional methods of video analysis. Our system offers many functionalities at different granularity that can be applied to applications with different requirements. The Toolbox also provides a flexible system for navigation and display using the low-level features or their combinations. Finally, the feature extraction according to the Toolbox can be performed in the compressed domain and preferably real-time system performance such as the videosurveillance systems.

3. Moving Object Detection And Extraction

To detect and extract a moving object from a video dataset we use a region-based active contours model where the designed objective function is composed of a region-based term and optimize the curve position with respect to motion and intensity properties. The main novelty of our approach is that we deal with the motion estimation by optical flow computation and the tracking problem simultaneously. Besides, the active contours model is implemented using a level set, inspired from Chan and Vese approach [3], where topological changes are naturally handled.

3.1 Motion estimation by optical flow

In our system, we use gradient-based optical flow algorithm proposed by Horn and Schunck [2]. similar to T. Macan and S. Loncaric [14],we have integrated the algorithm in multigrid technique where the image is decomposed into Gaussian pyramid-set of the reduced images. The calculation starts at a coarser scale of the image decomposition, and the results are propagated to finer scales.

Let us suppose that the intensity of the image at a time t and position (x, y) is given by I (x, y, t). The assumption on brightness constancy is made that the total derivative of brightness function is zero which results the following equation:

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t} = 0 \; or \; I_{x,i}u_i + I_{y,i}v_i + I_{t,i} = 0 \qquad (1)$$

This equation is named 'Brightness Change Constraint Equation'. Where u and v are components of optical flow in horizontal and vertical directions, respectively, and I_x, I_y and It are partial derivatives of I with respect to x, y and t respectively. Horn and Schunck added additional smoothness constraint because the equation (1) is insufficient to computem both components of optical flow. They minimized weighted sum of smoothness term and brightness constraint term:

$$\int_\Omega \left(I_x u + I_y v + It\right)^2 + \lambda\left(\|\nabla u\|^2 + \|\nabla v\|^2\right)dx \qquad (2)$$

Minimization and discretization of equation (2) results in two equations for each image point where vector values ui and vi are optical flow variables to be determined. To solve this system of differential equations, we use the iterative Gauss-Seidel relaxation method (for more detail see *http ://benallal.free.fr/an/Optim6/Optim6.htm*).

3.2 Our moving object segmentation model

In our case, taking into consideration the motion information obtained by calculating the optical flow, we propose the following descriptors for the segmentation of mobile objects in a video surveillance dataset:

$$k_{in}(x,\Omega_{in}) = \lambda\left|SV_g(x) - c_1(\Omega_{in})\right|^2$$
$$k_{out}(x,\Omega_{out}) = \lambda\left|SV_g(x) - c_2(\Omega_{out})\right|^2 \qquad (3)$$
$$k_b(x) = \mu$$

With c_1 is the average of the region Ω_{in}, c_2 is the average of the region Ω_{out}, μ and λ constants positive. $SV_g(x)$ is the image obtained after a threshold of the optical flow velocity and applicate of a gaussian filter. The values of c_1 and c_2 are re-estimated during the spread of the curve. The method of levels sets is used directly representing the curve $\Gamma(x)$ as the curve of zero to a continuous function U(x). Regions and contour are expressed as follows:

$$\Gamma = \partial\Omega_{in} = \{x \in \Omega_I \,/U(x)=0\}$$
$$\Omega_{in} = \{x \in \Omega_I \,/U(x)<0\} \qquad (4)$$
$$\Omega_{Out} = \{x \in \Omega_I \,/U(x)>0\}$$

The unknown sought minimizing the criterion becomes the function U. We introduce also the Heaviside function H and the measure of Diract δ_0 defined by:

$$H(z) = \begin{matrix} 1 \, if \, z \leq 0 \\ 0 \, if \, z > 0 \end{matrix} \quad et \; \partial_0(z) = \tfrac{d}{dz}H(z)$$

The criterion is then expressed through the functions U, H and δ in the following manner:

$$J(U,c_1,c_2) = \int_{\Omega_I} \lambda |SV_g(x)-c_1|^2 H(U(x)dx +$$
$$\int_{\Omega_I} \lambda |SV_g(x)-c_2|^2 (1-H(U(x)))dx + \int_{\Omega_I} \mu \delta U(x)) |\nabla U(x)| dx \qquad (5)$$

with:

$$c_1 = \frac{\int_\Omega SV_g(x)H(U(x))dx}{\int_\Omega H(U(x))dx}$$

$$c_2 = \frac{\int_\Omega SV_g(x)(1-H(U(x)))dx}{\int_\Omega (1-H(U(x)))dx} \qquad (6)$$

To calculate the Euler-Lagrange equation for unknown function U, we consider a regularized versions for the functions H and δ noted $H_\in$ and $\delta_\in$. The evolution equation is found then expressed directly with U, the function of the level set:

$$\frac{\partial U}{\partial T} = \delta_\in(U)[\mu div(\frac{\nabla U}{|\nabla U|}) + \lambda |SV_g(x)-c_1|^2 - \lambda |SV_g(x)-c_2|^2](in\Omega_I) \qquad (7)$$

$$\frac{\delta_\in(U)}{|\nabla U|}\frac{\partial U}{dN} = 0(on\partial\Omega_I)$$

with $div(\frac{\nabla U}{|\nabla U|})$ the curvature of the level curve of U viax and $\frac{\partial U}{dN}$ the derivative of U compared to normal inside the curve N.

3.3 Supported video surveillance concepts and events
Until now, our system supports 5 concepts and 6 events.

The 5 concepts supported by our system are as follows:

- C1: Approaching vehicule to the camera (figure 3.a)
- C2: One or more moving vehicule (figure 3.b)
- C3: Approaching pedestrian (figure 3.c)
- C4: One or more moving pedestrian (figure 3.d)
- C5: Combinated Concept (figure 3.e)

In our system we target six class of event. We divide this list into two categories:

Collaborative events:

- Embrace
- People Split Up

Individual events:

- Elevator No Entry
- Object Put
- Person Runs
- Opposing Flow

(a)　　　　　　　　　　(b)　　　　　　　　　　(c)

(d)　　　　　　(e)

Figure 3. Examples of images extracted from our video surveillance dataset

Figure 4. Elevator No entry event

Person run

Embrace

Figure 5. Person Run and Embrace events

4. Combining single SVM Classifier For Learning Video Event

Support Vector Machines (SVMs) have been applied successfully to solve many problems of classification and regression. However, SVMs suffer from a phenomenon called 'catastrophic forgetting', which involves loss of information learned in the presence of new training data. Learn++ [12] has recently been introduced as an incremental learning algorithm. The strength of Learn++ is its ability to learn new data without forgetting prior knowledge and without requiring access to any data already seen, even if new data introduce new classes. To benefit from the speed of SVMs and the ability of incremental learning of Learn++, we propose to use a set of trained classifiers with SVMs based on Learn++ inspired from [16]. Experimental results of detection of events suggest that the proposed combination is promising. According to the data, the performance of SVMs is similar or even superior to that of a neural network or a Gaussian mixture model.

(Opposing Flow)

People Split Up

Figure 6. Opposing Flow and People Split Up events

4.1 SVM Classifier

Support Vector Machines (SVMs) are a set of supervised learning techniques to solve problems of discrimination and regression. The SVM is a generalization of linear classifiers.The SVMs have been applied to many fields (bio-informatics, information retrieval, computer vision, finance ...).

According to the data, the performance of SVMs is similar or even superior to that of a neural network or a Gaussian mixture model. They directly implement the principle of structural risk minimization [15] and work by mapping the training points into a high dimensional feature space, where a separating hyperplane (w, b) is found by maximizing the distance from the closest data points (boundary-optimization). Given a set of training samples $S = \{(xi, yi)|i = 1, ..,m\}$, where $x_i \in R_n$ are input patterns, $y_i \in +1,-1$ are class labels for a 2-class problem, SVMs attempt to find a classifier $h(x)$, which minimizes the expected misclassification rate. A linear classifier $h(x)$ is a hyperplane, and can be represented as $h(x) = sign(w^T x+b)$. The optimal SVM classifier can then be found by solving a convex quadratic optimization problem:

$$\max_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \, subject \, to$$

$$y_i\left(\langle w, x_i \rangle + b\right) \geq 1 - \xi_i \, and \, \xi_i \geq 0 \tag{8}$$

Where b is the bias, w is weight vector, and C is the regularization parameter, used to balance the classifier's complexity and classification accuracy on the training set S. Simply replacing the involved vector inner-product with a non-linear kernel function converts linear SVM into a more flexible non-linear classifier, which is the essence of the famous kernel trick. In this case, the quadratic problem is generally solved through its dual formulation:

$$L(w,b,\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}(\sum_{i=1}^{m}y_i y_j \alpha_i \alpha_j \mathrm{K}(x_i,x_j))$$

$$subject \, to \, C \geq \alpha_i \geq 0 \, and \, \sum_{i=1}^{m} y_i \alpha_i y_i = 0 \tag{9}$$

where a_i are the coefficients that are maximized by Lagrangian. For training samples x_i, for which the functional margin is one (and hence lie closest to the hyperplane), $\alpha_i \succ 0$ Only these instances are involved in the weight vector, and hence are called the support vectors [6]. The non-linear SVM classification function (optimum separating hyperplane) is then formulated in terms of these kernels as:

$$h(x) = sign\left(\sum_{i=1}^{m}\alpha_i y_i K(x_i,x_j) - b\right) \tag{10}$$

4.2 M-SVM Classifiers

M-SVM is based on Learn++ algorithm. This latter, generates a number of weak classifiers from a data set with known label. Depending on the errors of the classifier generated low, the algorithm modifies the distribution of elements in the subset according to strengthen the presence of the most difficult to classify. This procedure is then repeated with a different set of data from the same dataset and new classifiers are generated. By combining their outputs according to the scheme of majority voting Littlestone we obtain the final classification rule.

The weak classifiers are classifiers that provide a rough estimate - about 50% or more correct classification - a rule of decision because they must be very quick to generate. A strong classifier from the majority of his time training to refine his decision criteria. Finding a weak classifier is not a trivial problem and the complexity of the task increases with the number of different classes, however, the use of NN algorithms can correctly resolved effectively circumvent the problem. The error is calculated by the equation:

$$error_t = \sum_{i:h_i(x_i)\neq y_i} S_t(i)\left[\left|h_t(x_i) \neq y_i\right|\right] \tag{11}$$

with $h_t : X \rightarrow Y$ an hypothesis and where TR_t is the subset of training subset and the TE_t is the test subset. The synaptic coefficients are updated using the following equation:

$$\tag{12}$$

Where t is the iteration number, B_t composite error and standard composite hypothesis H_t.
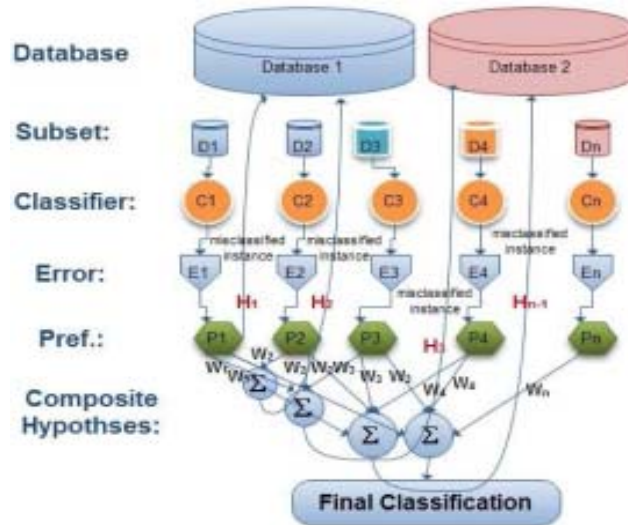

Figure 7. M-SVM classifier

In our approach we replace each weak classifier by SVM. After T_k classifiers are generated for each D_k, the final ensemble of SVMs is obtained by the weighted majority of all composite SVMs:

$$H_{final}(x) = arg\,\underset{y \in Y}{max} \sum_{k=1}^{K} \sum_{t:h_t(x)=y} log\frac{1}{\beta_t} \tag{13}$$

5. Visual Feature Extraction

We use a set of different visual descriptors at various granularities for each frame, rid of the static background, of the video shots. The relative performance of the specific features within a given feature modality is shown to be consistent across all concepts/events. However, the relative importance of one feature modality vs. another may change from one concept/event to the other. The following descriptors had the top overall performance for both search and concept modeling experiments:

• Color Histogram: global color represented as 128- dimensional histogram in HSV color space.

• Color Moments: localized color extracted from 3x3 grid and represented by the first 3 moments for each grid region in Lab color space as normalized 255-dimensional vector.

• Co-occurence Texture: global texture represented as a normalized 96-dimentional vector of entropy, energy, contrast and homogeneity extracted from the image grayscale co-occurence matrix at 24 orientation.

• Gabor Texture: Gabor functions are Gaussians modulated by complex sinisoids. The Gabor filter masks can be considred as orientation and scale-tunable and line detectors. The statistics of these micro-features in a given region can be used to characterize the underlying texture information. We take 4 scales and 6 orientations of Gabor textures and further use their mean and standard deviation to represent the whole frame and result in 48 textures.

• Fourier: Features based on the Fourier transform of the binarized edge image. The 2- dimensional amplitude spectrum is smoothed and down-sampled to form a feature vector of 512 parameters.

• Sift:The SIFT descriptor [9] is consistently among the best performing interest region descriptors. SIFT describes the local shape of the interest region using edge histograms. To make the descriptor invariant, while retaining some positional information, the interest region is divided into a 4x4 grid and every sector has its own edge direction histogram (8 bins). The grid is aligned with the dominant direction of the edges in the interest region to make the descriptor rotation invariant.

• Combined Sift and Gabor.

• Wavelet Transform for texture descriptor: Wavelets are hybrids that are waves within a region of the image, but otherwise particles. Another important distinction is between particles that have place tokens and those that do not. Although all particles have places in the image, it does not follow these places will be represented by tokens in feature space. It is entirely feasible to describe some images as a set of particles, of unknown position. Something like this happens in many description of texture. We performe 3 levels of a Daubechies wavelet [5] decomposition for each frame and calculate the energy level for each scale, which resulted in 10 bins features data.

• Hough Transform: As descriptor of shape we employ a histogram based on the calculation of Hough transform [10]. This histogram gives information better than those given by the edge histogram. We obtain a combination of behavior of the pixels in the image along the straight lines.

• Motion Activity: We use the information calculated by the optical flow, through concentrating on movements of the various objects (people or vehicle) detected by the method described in the previous section. The descriptors that we use are correspond to the energie calculated on every sub-band, by a decomposition in wavelet of the optical flow estimated between every image of the sequence. We obtain a vector of 10 bins, they represent for every image a measure of activity sensitive to the amplitude, the scale and the orientation of the movements in the shot.

6. Experimental Results

Experiments are conducted on the many sequence from TRECVid'2009 database of video surveillance and many other sequences from road traffics. About 20 hours are used to train the feature extraction system, that are segmented in the shots. These shots were annotated with items in a list of 5 concepts and 6 events. We use about 20 hours for the evaluation purpose.

To evaluate the performance of our concept detection subsystem, we use the common measure from the information retrieval community: the Average Precision. Figure 8 shows the evaluation of returned shots. The best results are obtained for concepts: 1,2,3,5. The remaining run also provide satisfying results.

	Run 1		Run 2	
Event	**A.NDCR**	**M.NDCR**	**A.NDCR**	**M.NDCR**
Embrace	1.045	0.981	1.025	0.961
PeopleSplitUp	3.432	0.983	3.331	0.953
ElevatorNoEntry	0.341	0.332	0.331	0.322
ObjectPut	1.076	0.997	1.056	0.952
PersonRuns	1.087	0.976	1.016	0.956
OpposingFlow	1.032	1.021	0.993	0.923

Table 1. Event detection Results

To evaluate the performance of event detection sub-system we use the TRECVID'2009 event detection metrics. The evaluation uses the Normalized Detection Cost Rate (NDCR). NDCR is a weighted linear combination of the system's Missed Detection Probability and False Alarm Rate (measured per unit time). The measure's derivation can be found in (http : //www.itl.nist.gov/ iad/mig/tests/trecvid/2009 /doc/EventDet09 " EvalPlan " v03.htm) and the final formula is summarized below. Two versions of the NDCR will be calculated for the system: the Actual NDCR and the Minimum NDCR.

The actual and minimum NDCRs for each of the events can be seen in Table 1. We have achieved very competitive minimum DCR results on the events of embrace, people Split UP, Object Put, opposing Flow and especially for Elevator No Entry. We did not extensively tune parameters with the aim of producing low actual DCR score; our actual DCR looks relatively higher (the lower the score, the better the performance). But our system achieved very good minimum DCR scores.

7. Conclusion

In this paper, we have presented a new approach for high-level feature extraction for video surveillance indexing and retrieval.

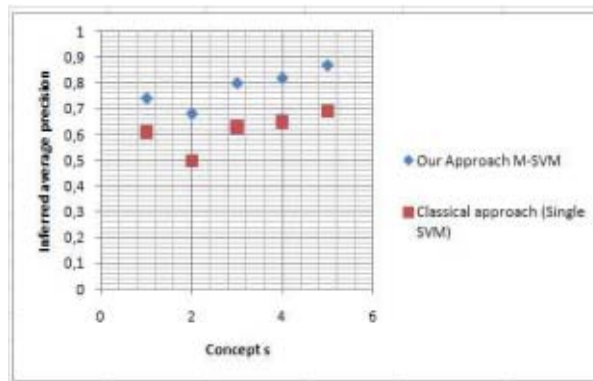Figure 8. Our run score versus Classical System (Single SVM) by Concept

The results obtained so far are interesting and promoters.The advantage of our approach is that allows human operators to use context-based queries and the response to these queries is much faster. The meta-data layer allows the extraction of the motion and objects descriptors from video key-frames to XML files that then can be used by external applications such as multimedia data mining systems. Finally, the system functionalities will be enhanced by a complementary tools to improve the basic concepts and events taken care of by our system.

8. Acknowledgement

References

[1] Wali, A., Alimi A, M.,(2009) Event detection from video surveillance data based on optical flow histogram and high-level feature extraction. *IEEE DEXA Workshops 2009*, pages 221–225.

[2] Horn, B, K, P., and Schunk, B,G.,(1981) Determining optical flow. *Artificial Intelligence*, 17:185–20.

[3] Tony Chan and Luminita Vese.,(1999) "An active contour model without edges. In *Scale-Space Theories in Computer Vision", Springer Berlin / Heidelberg, V. 1682*.

[4] De Rooij K, E, A., van de Sande, F, J., Seinstra , W, M., Smeulders, A, H, C., Thean, C, J., Veenman, D, C, Koelma., Van Liempt , A., and Worring, M.,(2006) The mediamill trecvid 2006 semantic video search engine. In *Proceedings of the 4th TRECVID Workshop*, Gaithersburg, USA, November.

[5] Daubechies, I.,(1991) *CBMS-NSF series in app. Math.*, chapter SIAM.

[6] Feki, I., Elleuch, N.,(2009) Regim at trecvid2009: Semantic access to multimedia data. In *TREC Video Retrieval Evaluation Online Proceedings*, Workshop of TRECVID.

[7] Ellouze M. Karray H. and Alimi M. A. Genetic algorithm for summarizing news stories. In *Proceedings of international conference on computer vision theory and applications*, pages 303–308, Spain, Barcelona, March 2006.

[8] Wali A. Karray H. and Alimi M.A. Sirpvct: System of indexing and the search for video plans by the contents text. In *Proc. Treatment and Analyzes information: Methods and Applications , TAIMA07*, pages 291–297, Tunisia, Hammamet, May 2007.

[9] Lowe David, G.,(2004) Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2(60):91–110.

[10] Boujemaaa, N., Ferecatu, M., and Gouet, V.,(2002) Approximate search vs. precise search by visual content in cultural heritage image databases. In *Proc. of the 4-th International Workshop onMultimedia Information Retrieval (MIR 2002) in conjunction with ACMMultimedia*.

[11] Karray, H., Ellouze, M., and Alimi, M,A., (2007)Using text transcriptions for summarizing arabic news video. In *Proc. Information and Communication Technologies International Symposuim , ICTIS07*, pages 324–328, Morocco, Fes,Davi.

[12] S. S. Udpa R. Polikar, L. Udpa and V. Honavar. Learn++: An incremental learning algorithm for supervised neural networks. *IEEE Trans. Sys. Man, Cybernetics (C*, 31(4):497–508, 2001.

[13] A. Yanagawa S. Chang, W. Jiang and E. Zavesky. Colombia university trecvid2007: High-level feature extraction. In *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID07, 2007.

[14] Macan T. and Loncaric S. Hybrid optical flow and segmentation technique for lv motion detection. *Proceedings of SPIE Medical Imaging, San Diego, USA*, pages 475–482, 2001.

[15] V. Vapnik. *Statistical Learning Theory*. 1998.

[16] Robi P. Zeki E. and al. Ensemble of svms for incremental learning. *LNCS MCS*, 3541:246–256, 2005.