REGIM at TRECVID2008: High-level Features Extraction and Video Search

Hichem Karray, Ali Wali, Nizar Elleuch, Anis Ben Ammar
Mahdi Ellouze, Issam Feki, Adel M. Alimi
REGIM: Research Group in Intelligent Machines, ENIS , Tunisia
[hichem.karray, ali.wali, nizar.elleuch,anis.benammar,mahdi.ellouze,issam.feki,adel.alimi]@ieee.org,
http://www.regim.org

Abstract

In this paper, we present an overview of a software platform that has been developed within the REGIMVid project for TRECVid 2008 evaluation. The platform termed REGIMVidToolbox provides Hight-level feature extraction from audio-visual content and video search. The toolbox is based on the MPEG-7 eXperimental Model (XM), with extensions to provide descriptor extraction from arbitrarily shaped image segments. Thereby, it supports local descriptors reflecting real image content. We describe the architecture of the toolbox as well as providing an overview of the descriptors supported to date. We also briefly describe the search task. We then demonstrate the usefulness of the toolbox in the context of feature extraction, concepts learning and retrieval in large collections of video dataset.

Keywords: *Feature fusion,classifier fusion, optical flow, Support Vector Machines.*

1 Introduction

Image and video indexing and retrieval continues to be an extremely active area within the broader multimedia research community [4, 12]. 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 (fig.1). 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 pre-processing is necessary in order to extract descriptors that can be subsequently linked to the concept and used in the reasoning process. In

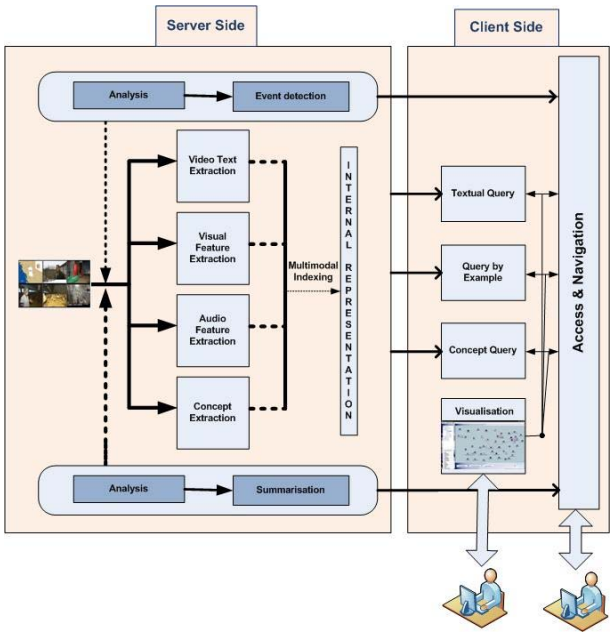

Figure 1: REGIMVid platform Architecture

addition to concept-based reasoning, the project has other research activities that require high-level feature extraction (e.g. semantic summary of metadata [6], Text-based video retrieval [10, 7]) it was decided to develop a common platform for descriptor extraction that could be used throughout the project. The requirements for the platform for High-level feature extraction were that it should:

- Provide extraction of a subset of MPEG-7 features;
- Facilitate the integration of new descriptors;
- Facilitate extraction of global (i.e. corresponding to the entire image) and local (i.e. corresponding to sections of the entire image only) descriptors;
- Be independent and suitable for integration into larger scale demonstration systems.

In this paper we describe the version of the platform developed within the framework of our participation in TRECVid 2008 [13]. The remainder of the paper is organised as follows: a general overview of the toolbox is provided in Section 2. It includes a description of the architecture, the descriptors of the visual feature extraction and the TRECVid 2008 Feature extraction results. In Section 3, we present our participation in the search task. Finally, we describe our future plans for both the extension of the toolbox and its use in different scenarios.

2 REGIMVidTOOLBOX OVERVIEW

In this section, we present an overview of the structure of the toolbox and briefly describe the audio and visual feature extraction techniques currently supported. The REGIMVidToolbox currently supports extraction of 20 low-level audio-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.

2.1 Architecture

The main objectif of our system is to provide automatic content analysis using frame-based and low-level features. The system (fig.2), first extracts features at the frame level and then labels each frame based on corresponding features. For example, if three features are used (color, motion and audio), each frame has at least three labels.

This reduces the video to labels sequences sets, there being one sequence of labels for feature common among consecutive frames. The multiple label sequences retain considerable information, while simultaneously reducing the video into a simple form.

It should be apparent to those of ordinary skill in the art, that the amount of data required to code the labels is orders of magnitude less than the data that encodes the video itself. This simple form enables machine learning techniques such as Support Vector Machines to perform high-level feature extraction.

The procedures according to the system, offer a way to combine low-level features which enhances the system performance. The high-level feature extraction system according to the Toolbox provides an open framework that enables easy integration with new features. Furthermore, the Toolbox can be integrated with traditional methods of video analysis. Our system provides functionalities at different granularities that can be applied to applications with different requirements. The Toolbox also provides a system for

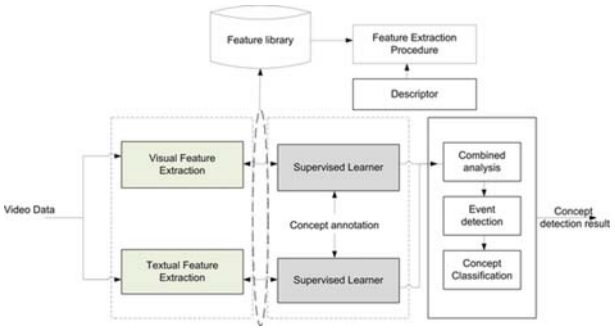

Figure 2: Overview of the REGIMVidToolbox for video input

flexible browsing or visualisation using individual low-level features or their combinations. Finally, the feature extraction according to the Toolbox can be performed in the compressed domain for fast, and preferably real-time, system performance.

2.2 Supervised Learners

We perceive concept detection in video as a pattern recognition problem. Given pattern x, part of a shot i, the aim is to obtain a probability measure, which indicates whether the corresponding semantic concept C_x is present in shot i. Similar Similar to the MediaMill system [3], REGIMVidToolbox uses the Support Vector Machines (SVM) framework for supervised learning of concepts. Here we use the LIBSVM implementation (for more detail see $http://www.csie.ntu.edu.tw$) with radial basis function and probabilistic output. We obtain good SVM parameter settings by using an iterative search on a large number of SVM parameter combinations.

The REGIMVidToolbox optimizes SVM parameters that aim to balance positive and negative examples. We measure average precision performance of all parameter combinations and select the combination that yields the best performance.

In addition to the SVM we also experiment with Hidden Markov Models (HMM) for the object Related concepts (fig.3). This classifier is known to be less effective than SVM, in terms of concept detection performance, they require no parameter tuning so classification is relatively cheap. All two classifiers yield a probability measure $p(C_j|x_i)$, which we use to rank and to combine concept detection results. We use a Hidden Markov Model (HMM) Toolbox written by Kevin Murphy (1998). (See http://www.ai.mit.edu/ murphyk/Software/hmm.html for details).

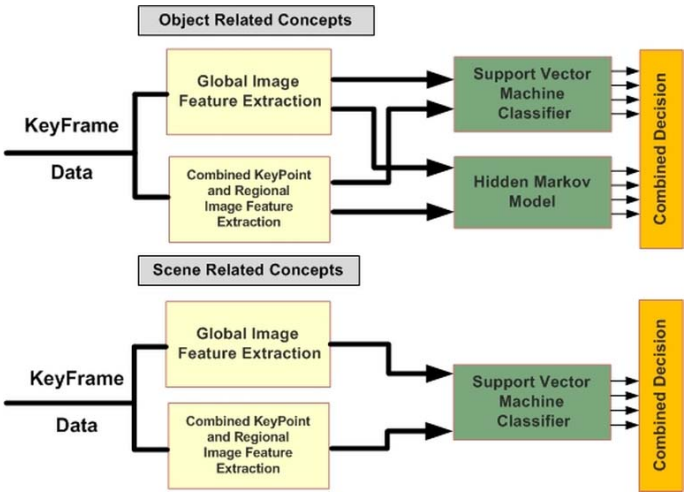

Figure 3: Simplified Overview of the REGIMVidToolbox for the TRECVid'2008 participation.

2.3 Visual Feature Extraction

We used a set of different vsual descriptors at various granularities for each representative keyframe of the video shots. The relative performance of the specific features within a given feature modality is shown to be consistent across all concepts/topics. However, the relative importance of one feature modality vs. another may change from one concept 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.
- Edge Histogram: global edge histograms with 8 edge direction bins and 8 edge magnitude bins, based on a sobel filter (64-dimensional).
- Co-occurence Texture: global texture represented as a normalized 96-dimentional vector of entropy, energy, contrast and homogeneity extracted from the image gray-scale 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 keyframe and result in 48 textures.
- GLCM: The GLCM (Gray-level co-occurrence matrix) is a common technique in statistical image analysis that is used to estimate image properties related to second-order statistics. GLCM considers the relation between two neighboring pixels in one offset, as the second order texture, where the first pixel is called reference and the second one the neighbor pixel. GLCM is the two dimensional matrix of joint probabilities $P_{d,\theta}(i,j)$ between pairs of pixels, separated by a distance d in a given direction θ [2]. We used a 4 statistical features (Contrast, Correlation, Energy and Homogeneity) from gray-level co-occurrence matrix for texture classification.
- Fourier: The Fourier-transforming said image to find out a radial reference point, normalizing said Fourier-transformed image with reference to said reference point, and then describing said texture descriptor by using said normalized values of said Fourier-transformed image. Here, the radial reference point is set by determining an arc in which one of energy, entropy and a periodical component of said Fourier-transformed image apart at the same distance from the origin in said frequency domain is most distributed, and then setting a radius of said founded arc as said radial reference point.
- Sift:The SIFT descriptor [8] 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 that 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 performed 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 [9]. This histogram gives information better than those given by the edge histogram. We thus obtain a combination of behavior of the pixels in the image along the straight lines.

2.4 TRECVid'2008 Evaluation Results

Experiments are conducted on the TRECVid 2008 database of news magazine, science news, news reports, documentaries, educational programs, and archival video. About 100 hours are used to train the feature extraction system, that are segmented in the shots. These shots where annotated with items in a list of 20 labels and 100 hours are used for the evaluation purpose. The training set is divided into two subsets in order to train classifiers and subsequently the fusion parameters. For evaluation, we used the common measure from the information retrival community: the Average Precision. Figure 4 shows the

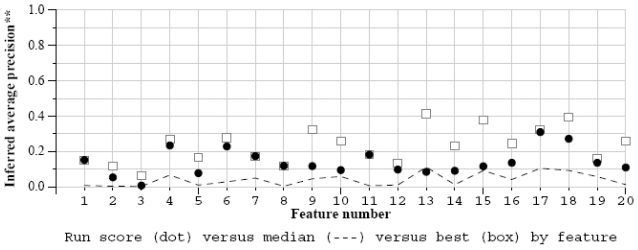

Figure 4: TRECVID 2008: Feature extraction results.

evaluation of returned shots. The best results were obtained for concepts: 1, 7, 8 and 11. The remaining runs also provides satisfying results especially for the 4, 12, 17 and 19 concepts.

3 Search

3.1 Our Search System Overview

The REGIM Group participated in the search task for TRECVid 2008 and submitted one run based on interactive search (fig.5). We describe this run below.

3.2 Face detection

To detect the faces in the images resulting from the video sequences, we used a method based on neural network [14]. The system is composed of a face detection stage based on skin color and a neural network. The resulting system is very efficient to find specific face images and to cope with the different face conditions in a video sequence.

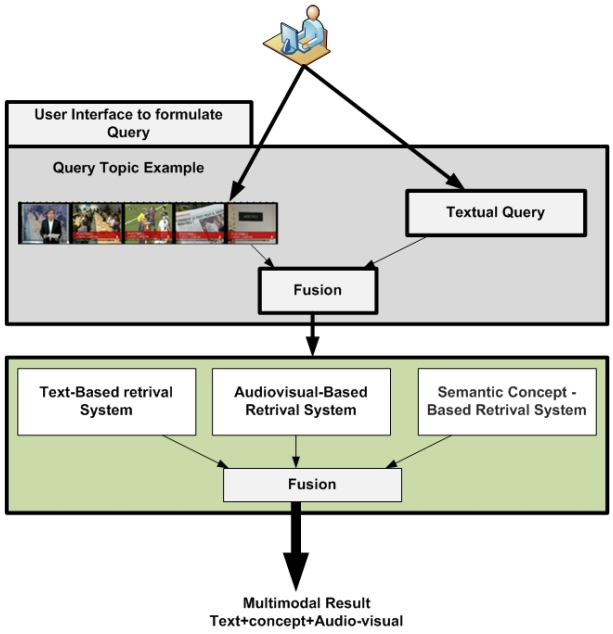

Figure 5: Overview of REGIM Interactive Search System.

3.3 Vehicule moving/approaching detection

To detect a moving or approaching vehicule we use an approach developped in REGIM Group for image segmentation by motion based on optical flow histogram with no prior knowledge of the motion nature [1]. First, we start by estimating the motion from images sequence using optical flow technique. Second, we perform a classification using the histogram of the optical flow vectors to distinguish between the objects which approaches the camera of those which are moving away from the camera. Finally we use a chain coding algorithm that we applied to each class for the spatial segmentation.

3.4 One or more peopel detection

To detect one or more peopel in several images, we use a non-parametric thresholding method that we term Mutual-Information Thresholding developped by C. O'Conaire[11]. The system worked as follows. First, the median background image was computed. Then for each image, two detectors were used: one based on pedestrian contour and the other based on silhouette. The contour detection map was obtained by convolving the pedestrian contour template with the Sobel edges of the image. The silhouette detection map was obtained by convolving the pedestrian silhouette template with the absolute difference image between the current image and the background image. Thresholds for these maps were obtained using a mutual information thresholding algorithm. Pedestrian regions were determined as all pixels that had above thresh-

old values in both maps. Next, each local maxima in the contour detection map within these regions was paired with the closest local maxima in the silhouette detection map within these regions. Maxima in the silhouette detection map were then paired with the closest maxima in the contour detection map. Person candidates corresponded to each pair of maxima, from the two separate maps, that were both paired to each other.

3.5 Video-Text detection and Extraction

Text within an image is of particular interest for indexing and retrieval of video because of its capability to describe the contents of an image, and its relationship to the semantic information. It also enables applications such as keyword based search in multimedia databases. In this work we use a wavelet-based method to detect and extract the text from the video frames [7]. This method works without the dependency on the availability of an OCR. Extraction of text information involves detection, localization, enhancement and recognition of the textual content in the video frames. A method involves a frame by frame processing on the entire video for locating textual blocks.

3.6 Evaluation Results

This section describes results of a run that we submitted to the TRECVID2008 search task. 24 (221–244) Test topics evaluated by NIST are used in the following evaluation. Figure 6 shows inferred average precisions obtained with the topics-based video retrival where the topic example is introduced by the user. These results prove that our method gives a better precision. Figure 7 present the elapsed search time (mins) by topic.

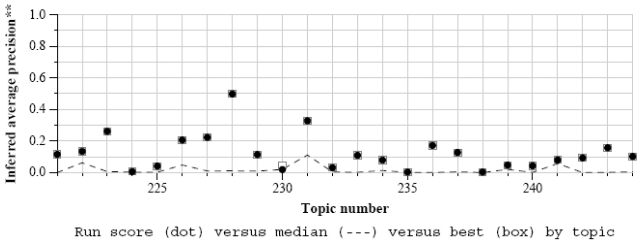

Figure 6: TRECVID 2008: Search results.

4 Conclusion

REGIM Research Group participated in the TREC Video Retrival High-level features extraction and Search tasks for the first time. In this paper, we have presented preliminary results and experiments

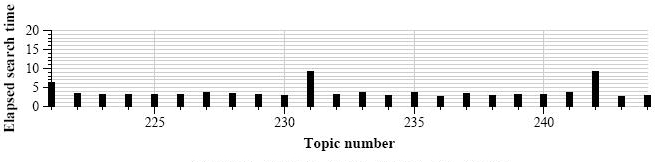

Figure 7: TRECVID 2008: elapsed search time (mins) by topic.

for both tasks. The REGIMVidToolbox functionalities will be enhanced by a complementary tools as personalization and visualization. These last subsystems are under development.

References

[1] Wali A. and Alimi M. A. Mosmof: Moving object segmentation by motion based on optical flow histogram. In *proceeding of IEEE International Symposium on Image and Video Communication(ISIVC08)*, Spain, Bilbao, July 2008.

[2] Miroslav B. and Robert H. *Novel Method for Color Textures Features Extraction Based on GLCM*, chapter RADIOENGINEERING. 2007.

[3] J. C. van Gemert J.-M. Geusebroek C. G. M. Snoek, M.Worring and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the ACM International Conference on Multimedia*, pages 421–430, Santa Barbara, USA, October 2006.

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

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

[6] 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.

[7] 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.

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

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

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

[11] Eddie C. O'Conaire, C. O' Connor N. and Alan S. *Detection Thresholding Using Mutual Information.* PhD thesis, http://elm.eeng.dcu.ie/oconaire/papers/, May 2008.

[12] 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.

[13] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330. ACM Press, 2006.

[14] T. Wittman. Face detection and neural networks. Technical report, Department of Mathematics, University of Minnesota, December 2001.