# On the selection of MPEG-7 Visual Descriptors and their Level of Detail for Nature Disaster Video Sequences Classification

Javier Molina[1], Evaggelos Spyrou[2], Natasa Sofou[2], José M. Martínez[1]

[1]Grupo de Tratamiento de Imágenes, Universidad Autónoma de Madrid, Spain
[2]Image, Video and Multimedia Laboratory, National Technical University of Athens, Greece
Javier.Molina@uam.es, espyrou@image.ece.ntua.gr, natasa@image.ntua.gr , JoseM.Martinez@uam.es

Abstract. In this paper, we present a study on the discrimination capabilities of colour, texture and shape MPEG-7 [1] visual descriptors, within the context of video sequences. The target is to facilitate the recognition of certain visual cues which would then allow the classification of natural disaster-related concepts. Low-level visual features are extracted using the MPEG-7 “eXperimentation Module” (XM) [2]. The extraction times associated to the levels of detail of the descriptors are measured. The pattern sets obtained as combination of significant levels of detail of different descriptors are the input to a Support Vector Machine (SVM), resulting on the classification accuracies. Preliminary results indicate that this approach could be useful for the implementation of real-time spatial regions classifiers.

Keywords: Image Classification, Semantic Retrieval, Visual Descriptors

## 1 Introduction

Due to the huge amount of multimedia contents, the automatic extraction of visual information from videos is widely required. Time performance evaluation is gaining more and more importance in visual descriptor based applications, although not very much work has been done on this line [3]. Studies mainly focus on evaluation of images retrieval accuracies as in [4]. We propose a methodology that by operating on image regions, studies the relation between the classification accuracy and the computational cost of the required extractions of descriptors.

## 2 Visual Descriptors Profiles and their Levels Of Detail Selection

In this paper the term *Profile* refers to a combination of visual descriptors and the term *Levels of detail* refer to the different combinations of *elements* of a specific descriptor. Since this work gives focus to spatial regions of still images, it appears that some visual descriptors do not make sense in the presented application field. First of all, motion descriptors are discarded, because currently we are working at the frame-by-frame level. The *Group-of-Frame/Group-of-Picture Descriptor* is discarded since it is used for joint representation of a group of frames or pictures. Moreover, the *3-D Shape Descriptor* is directly ruled out from this study, simply because we are working with 2-D projections of the 3-D real world. The *Edge Histogram Descriptor* is only applicable to rectangular images, and not to arbitrary shape regions, thus, its use makes sense for global extraction from an image. Towards the goal of obtaining useful information for real-time implementations, it makes sense to rule out the *Texture Browsing Descriptor,* since its extraction (with the XM) is about 20 times slower than the *Homogeneous Texture Descriptor* as documented in [5]. Table 1 presents the extraction times for different levels of detail of each descriptor, grouping them (one label for each group) in the case they present low variation (Groups of levels of detail of a descriptor: mean extraction time ± percentage variation). As a result, each group of levels of detail gets represented by the most detailed descriptor.

Table 1. Levels of detail Extraction times.VP(Variances Present); BN (Bins Number); SC (Spatial Coherency); NBPD (Number of Bit Planes Discarded); QR (Quantification Resolution). Extraction times measured while executing the XM version 6.1 [2] on an Intel(R) Core(TM) 2 Duo CPU T7200 @ 2Ghz with 1GB of RAM.

| Label | Descriptors and levels of detail[1] | Extraction time(msec) |
|---|---|---|
| DCD1 | DCD: VP (1,0), BN (256,128,64,32,16,8,4,2), SC (1) | $1832 \pm 0.50\%$ |
| DCD2 | DCD: VP (1,0), BN (256,128,64,32,16,8,4,2), SC (0) | $1577 \pm 1.10\%$ |
| SCD | SCD: NBPD(0,1,2,3,4,5,6,8), NC(256,128,64,32,16) | $196 \pm 0.05\%$ |
| CSD | CSD: QR(256,128,64,32) | $191 \pm 0.15\%$ |
| CLD | CLD | 65 |
| HTD | HTD: layer 1 or layer 0 | $2652 \pm 0.05\%$ |
| rSD | region-based SD | 1933 |
| cSD | contour-SD | 315 |

Using the equation (1) with values: $n_{DCD} = 2+1$, $n_{SCD} = 1+1$, $n_{CSD} = 1+1$, $n_{CLD} = 1+1$, $n_{HTD} = 1+1$, $n_{region-basedSD} = 1+1$ and $n_{contour-SD} = 1+1$ , the obtained number of patterns´ sets ( $Ns$ ) is 96. The '$+1$' is added in order to consider the non usage of descriptors.

$$Ns = n_{DCD} \times n_{SCD} \times n_{CSD} \times n_{CLD} \times n_{HTD} \times n_{\text{region-based SD}} \times n_{\text{contour-SD}} \quad (1)$$

Each pattern set is analyzed with a Support Vector Machine using the LIBSVM implementation [6] and applying 10-fold cross-validation. The relations of computational cost and classification accuracy are presented in section 3 .

## 3 Classification Results

The dataset is composed of a subset of MESH repository, images from the *Labelme* dataset [7] and various images collected from the *world wide web*. The goal was the

[1] DCD (*Dominant Color Descriptor*); SCD (*Scalable Color* Descriptor) ; CSD (*Color Structure Descriptor*); HTD (*Homogeneous Texture Descriptor*); CSD (*Color Layout Descriptor*); region-based SD (*Region-Based Shape Descriptor*); contour-SD (*Contour-Shape Descriptor*)

detection of certain visual cues (*flames, smoke, vegetation, buildings, water, snow* etc) which, when combined, would assist to the detection of semantic concepts (*forest fires,floods, volcanic eruptions* etc). Around 100 spatial regions of each visual cue have been manually obtained and annotated.

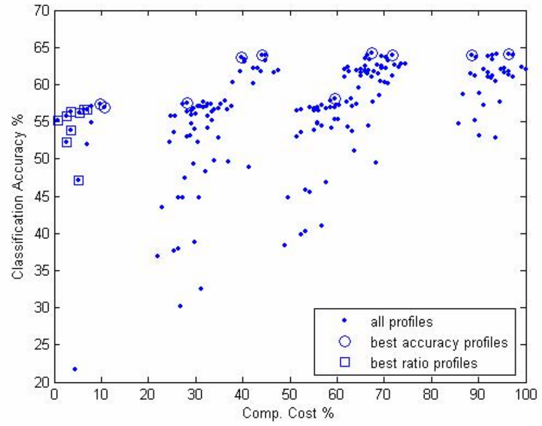

Fig. 1. Relation between computational cost and classification accuracy for the combinations of descriptors and their levels of detail. The computational cost is expressed as the percentage of the maximum value of computational cost (7184 msec.) which corresponds to the profile [DCD1 SCD CSD CLD HTD rSD cSD]

Table 2. Profiles with best classification accuracies of computational cost segments.

| Profile | Cost (%) | Acc. (%) | Acc./Cost |
|---|---|---|---|
| [SCD CSD cSD] | (0,10] | 57.41 | 5.8762 |
| [SCD CSD CLD cSD] | (20,30] | 56.98 | 5.3352 |
| [DC2 SCD CSD CLD] | (30,40] | 57.56 | 2.0382 |
| [SCD HTD] | (30,40] | 63.66 | 1.606 |
| [SCD HTD cSD] | (40,50] | 63.95 | 1.4524 |
| [DC2 SCD CSD CLD HTD] | (50,60] | 61.63 | 0.94583 |
| [SCD CLD HTD rSD] | (60,70] | 64.24 | 0.95227 |
| [SCD CLD HTD rSD cSD] | (70,80] | 63.95 | 0.89017 |
| [DC2 SCD HTD rSD] | (80,90] | 63.95 | 0.7226 |
| [DC1 SCD HTD rSD cSD] | (90,100] | 64.1 | 0.66466 |

Table 3. Profiles with best relation between accuracy and computational cost

| Profile | Cost (%) | Acc. (%) | Acc./Cost |
|---|---|---|---|
| [CLD] | 0.9 | 55.09 | 61.211 |
| [SCD] | 2.73 | 55.81 | 20.443 |
| [CSD] | 2.66 | 52.18 | 19.617 |
| [SCD CLD] | 3.63 | 56.4 | 15.537 |
| [CSD CLD] | 3.56 | 53.78 | 15.107 |
| [SCD CSD] | 5.39 | 56.25 | 10.436 |
| [SCD CSD CLD] | 6.29 | 56.69 | 9.0127 |
| [CLD cSD] | 5.29 | 47.09 | 8.9017 |
| [SCD cSD] | 7.11 | 56.69 | 7.9733 |
| [CSD cSD] | 7.04 | 51.89 | 7.3707 |

It can be inferred from the compiled data in Table 1 that in some combinations the use of a concrete descriptor is counterproductive, worsening the classification results. It makes sense to highlight the profiles which show the highest accuracy/computational cost ratios (Table 2). This gives us a hint of which MPEG-7 visual descriptors are most recommendable for real time contexts.

## 4 Conclusions & Future Work

In this work, a method for the estimation of the interdependency between classification accuracy of spatial regions and computational cost of the required descriptor extraction has been proposed. We can conclude that certain descriptors seem to be more recommendable for real time applications, since they improve the classification accuracy with a very small increase of the computational cost. Future work includes study on temporal dependant descriptors, such as motion or shape evolution descriptors.

**Acknowledgments.** The work presented in this paper was partially supported by the European Commision under contractFP6-027685 MESH, and by the Spanish Government under project TIN2004-07860-C02 (MEDUSA) and by the Comunidad de Madrid under project S-0505/ TIC-0223 (ProMultiDis-CM). Evaggelos Spyrou is funded by PENED 2003 – Project Ontomedia 03ED475.

## References

1 Manjunath, B.S.; Salembier, P.; Sikora, T.; “Introduction to MPEG-7”; 1st edition. John Wiley & Sons, Ltd.; West Sussex, England.

2 MPEG-7: Visual experimentation model (xm) version 10.0. ISO/IEC/JTC1/SC29/WG11, Doc. N4062 (2001).

3 Mikolajczyk, K. ,Schmid, C., “A performance Evaluation of Local Descriptors”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 10, 2007, pp.1615 – 1630.

4 Timo Ojala, Markus Aittola, Esa Matinmikko, "Empirical Evaluation of MPEG-7 XM Color Descriptors in Content-Based Retrieval of Semantic Image Categories," *icpr*, p. 21021, 16th International Conference on Pattern Recognition (ICPR'02) - Volume 2.

5 Ojala T, Mäenpää T, Viertola J, Kyllönen J & Pietikäinen M (2002) Empirical evaluation of MPEG-7 texture descriptors with a large-scale experiment. Proc. 2nd International Workshop on Texture Analysis and Synthesis, Copenhagen, Denmark, 99-102..

6 Chih-Chung Chang and Chih-Jen Lin, “LIBSVM : a library for support vector machines”, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

7 Russell, B.C., Torralba, A. Murphy, K.P. and Freeman, W.T. “LabelMe: a database and web-based tool for image annotation” MIT AI Lab Memo AIM-2005-025, 09/2005.