OPTIMAL COMBINATION OF LOW-LEVEL FEATURES FOR SURVEILLANCE OBJECT RETRIEVAL[*]

Virginia Fernandez Arguedas, Krishna Chandramouli, Qianni Zhang and Ebroul Izquierdo
Multimedia and Vision Research Group, School of Electronic Engineering and Computer Science
Queen Mary, University of London, Mile End Road, London, E1 4NS, U.K.

Keywords: Object retrieval, Multi-feature fusion, Particle swarm optimisation, Surveillance videos, MPEG-7 features, Machine learning.

Abstract: In this paper, a low-level multi-feature fusion based classifier is presented for studying the performance of an object retrieval method from surveillance videos. The proposed retrieval framework exploits the recent developments in evolutionary computation algorithm based on biologically inspired optimisation techniques. The multi-descriptor space is formed with a combination of four MPEG-7 visual features. The proposed approach has been evaluated against kernel machines for objects extracted from AVSS 2007 dataset.

1 INTRODUCTION

Recent technological developments coupled together with people's concern for safety and security have caused a wide spread application of Closed Circuit Television (CCTV) cameras which have been widely installed for surveillance monitoring. With such an exponential increase in video footage, there exists critical need for the development of automatic and intelligent retrieval models for objects and events to enable efficient media access, navigation and retrieval. Addressing the challenges related to object indexing, several approaches has been presented based on probabilistic, statistical and biologically inspired classifiers (Chandramouli and Izquierdo, 2010). Many of these techniques generate satisfactory results for general datasets such as movies, sports and news. However, the challenge of retrieving surveillance objects remains a largely an open issue.

Among the approaches presented in the literature, visual appearance based retrieval has gained much popularity. The range of visual features used for object retrieval from surveillance videos include, colour histograms from different colour space and Gabor filters. More recently, MPEG-7 based colour, texture and shape descriptors have been largely investigated for multimedia indexing and retrieval (Sikora, 2002). In many of these approaches authors consider a single low-level descriptor to provide a high-level degree of distinguishability among objects. In order to generate robust and complex representation of objects, a multi-descriptor feature space is constructed to represent objects extracted from surveillance videos (Mojsilovic, 2005). The combination of low-level-features to obtain higher order representations have been addressed over the years in pattern recognition. For instance, in (Zhang and Izquierdo, 2007; Soysal and Alatan, 2003) authors proposed approaches that used combination of multiple low-level features to index and retrieve media items. However, to the best of our knowledge, such feature fusion approaches has not yet been applied for object retrieval from surveillance video datasets.

In this paper, we present an optimal combination of low-level feature spaces appropriate for surveillance object retrieval. Besides, in order to study the performance of the proposed multi-feature space a comparison against the individual features performance along with a linear combination of selected features is presented. The proposed retrieval framework exploits the recent developments in evolutionary computational algorithms based on biologically inspired optimisation techniques. Recent developments in optimisation techniques have been inspired by problem solving abilities of biological organisms such as bird flocking and fish schooling. One such technique developed by Eberhart and Kennedy is called Particle Swarm Optimisation (PSO)(Kennedy and Eberhart, 2001). The proposed approach has been

[*] The research was partially supported by the European Commission under contract FP7-216444 PetaMedia.

Fernandez Arguedas V., Chandramouli K., Zhang Q. and Izquierdo E..
OPTIMAL COMBINATION OF LOW-LEVEL FEATURES FOR SURVEILLANCE OBJECT RETRIEVAL.
DOI: 10.5220/0003527101870192
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications* (SIGMAP-2011), pages 187-192
ISBN: 978-989-8425-72-0

evaluated against three kernel machines for objects extracted surveillance dataset. From the study of evaluation results, we note the improved performance of the proposed approach across all concepts as opposed to improved performance for a single concept. The dataset has been specifically designed to be noisy in order to measure the robustness of the proposed optimal combination of the low-level feature space.

The remainder of the paper is structured as follows. In Section 2, an overview of the proposed surveillance object retrieval framework is presented followed by a brief introduction of Particle Swarm classifier in Section 3. Section 4 outlines the contribution of optimally combining low-level visual descriptor space. The experimental results are discussed in Section 5, followed by conclusion and future work in Section 6.

2 SURVEILLANCE OBJECT RETRIEVAL FRAMEWORK

The proposed surveillance object retrieval framework is presented in Fig.1. The framework integrates a training phase, a feature extraction component and object retrieval module. The training phase consists of the multi-feature fusion algorithm, which is used to create visual models to enable optimal combination of multiple low-level features. The multi-feature fusion algorithm is discussed in detail in Section 4. In the feature extraction phase, the video is subjected to motion analysis component to extract the blobs from the surveillance videos. The motion analysis component is based on an adaptive background subtraction algorithm based on Stauffer and Grimson approach (Stauffer and Grimson, 2000), followed by a spatial segmentation based on connected components and a temporal segmentation performed by a linearly predictive multi-hypothesis tracker. Finally, the retrieval phase is based on the particle swarm classifier. The classifier is based on evolutionary computation models, simulating the effects of fish schooling and bird flocks. The classifier is implemented for the multi-descriptor feature space whose performance is influenced by the weights derived for non-linear optimal combination of low-level feature space. The outcome of the classifier is a ranked list of objects retrieved from the image database, which are further evaluated against ground truth.

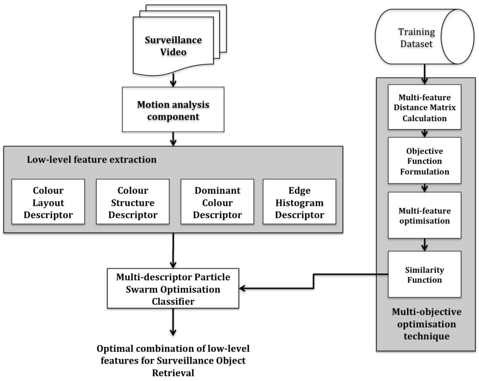
Surveillance Video
Motion analysis component
Low-level feature extraction
Colour Layout Descriptor
Colour Structure Descriptor
Dominant Colour Descriptor
Edge Histogram Descriptor
Multi-descriptor Particle Swarm Optimisation Classifier
Optimal combination of low-level features for Surveillance Object Retrieval
Training Dataset
Multi-feature Distance Matrix Calculation
Objective Function Formulation
Multi-feature optimisation
Similarity Function
Multi-objective optimisation technique

Figure 1: Framework overview.

Due to surveillance videos nature, a really time-consuming analysis processes a huge amount of information, where most of it belong to their quasi-static background proving no useful data. *Motion analysis component*'s objective is to improve the computational efficiency of the system and to provide movement information about the surveillance video objects. The three-step real-time *Motion Analysis Component* procures individual blobs to the *Low-level feature extraction Component*. Despite many advantages of the use of the *Motion Analysis Component*, object detection from surveillance videos is affected by several external factors as highlighted in Fig.2.

3 PARTICLE SWARM CLASSIFIER

In the PSO algorithm (Eberhart and Shi, 2001), the birds in a flock are symbolically represented as particles. A particle's location in the multidimensional problem space represents one solution for the problem. When a particle moves to a new location, a different solution to the problem is generated. The particles at each time step are considered to be moving towards particle's personal best (*pbest*) and swarm's global best (*gbest*). The motion is attributed to the velocity and position of each particle. Acceleration (or velocity) is weighted with individual parameters governing the acceleration being generated for c_1 and c_2. The equations governing the velocity and position of each particle are presented in Eq. 1 and 2.

$$\begin{aligned} v_{it}(t+1) &= v_{id}(t)+c_1(pbest_i(t)-x_{id}(t)) \\ &+ c_2(gbest_d(t)-x_{id}(t)) \qquad (1) \end{aligned}$$

$$x_{id}(t+1) = x_{id}(t)+v_{id}(t+1) \qquad (2)$$

where $v_{id(t)}$ represents the velocity of particle x in $d-$ dimension at time t, $pbest_i(t)$ represents the personal best solution of particle i at time t, $gbest_d(t)$ represents the global best solution for $d-$ dimension at time t, $x_{id}(t)$ represents the position of the particle

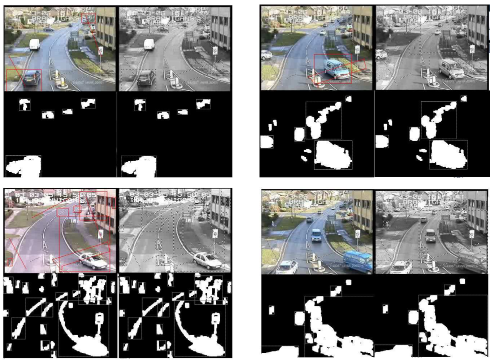

Figure 2: Motion analysis component results. Background subtraction and spatial segmentation techniques results can be observed for two different problematic situations as low quality image (top-left),videos with inaccurate background substration (top-right), videos with camera movement (bottom-left) and objects merged due to noise and shadow (bottom-right).

x in $d-$ dimension at time t and c_1, c_2 are constant parameters.

The first part of Eq. 1 represents the velocity at time t, which provides the necessary momentum for particles to move in the search space. During the initialisation process, the term is set to '0' to symbolise that the particles begin the search process from rest. The second part is known as the "cognitive component" and represents the personal memory of the individual particle. The third term in the equation is the "social component" of the swarm, which represents the collaborative effort of the particles in achieving the globally best solution. The social component always clusters the particles toward the global best solution determined at time t.

The PSO optimisation has been applied to improve the performance of Self Organising Maps (SOM), which is based on competetive learning scheme as discussed by Xu et al in (Xu and II, 2005). Briefly, the basic SOM training algorithm the input training vectors are trained with Eq. 3

$$m_n(t+1) = m_n(t) + g_{cn}(t)[x - m_n(t)], \tag{3}$$

where m is the weight of the neurons in the SOM network, $g_{cn}(t)$ is the neighbourhood function that is defined as in Eq. 4,

$$g_{cn}(t) = \alpha(t) exp(\frac{||r_c - r_i||^2}{2\alpha^2(t)}), \tag{4}$$

The PSO optimisation is achieved by evaluating the $L1$ norm between the input feature vector and the feature vector of the winner node. The global best solution obtained after the termination of the PSO algorithm is assigned as the feature vector of the winner node. The training process is repeated until all the input training patterns are exhausted. In the testing phase, the distance between the input feature vector is compared against the trained nodes of the network. The label associated with the node is assigned to the input feature vector.

4 MULTI-OBJECTIVE OPTIMISATION TECHNIQUE

In this paper, a fusion technique of multiple visual descriptors called *Multi-objective optimisation technique (MOO)* is presented. The objective is to learn associations between complex combinations of low level visual descriptors and the semantic concepts under study. As a result, the visual descriptors association is expected to complement each other improving their individual performance and overcoming their individual flaws. *MOO* aims to reduce the influence of noise coming from the background and identify an optimal mixture of visual descriptors to describe each semantic concept. In fact, the descriptors are combined according to a concept-specific metric, acquired during a training/learning stage from a set of representative blobs.

The challenge in *Multiple-objective optimisation technique (MOO)* is to find an optimal metric combining several low-level features and the suitable weights for such a combination. The *MOO* technique is a four-step process (Zhang and Izquierdo, 2007):

1. Distance Matrix Calculation. Four low-level features were extracted for each blob provided by the motion analysis. The provided training dataset is composed of as many entries as the number of training blobs, K, and four descriptors per blob. Considering all the entries of the dataset, composed of multiple descriptors. Foe each of such descriptor, a centroid is calculated generating a virtual centroid vector called $\bar{V} = (\bar{v}_{F_1}, \bar{v}_{F_2}, \bar{v}_{F_3}, \bar{v}_{F_4})$. Then, every distance between each blob low-level-feature descriptor and the respective centroid vector is calculated, obtaining the *multi-feature distance matrix*, *D*, which is the basis to build the objective functions for optimisation.

2. Objective Functions Formulation. In order to calculate an appropriated combined metric, a weighted linear combination of the feature descriptor distances (also called *objective function*) is proposed:

$$D^{(k)}(V^{(k)}, \bar{V}, A) = \sum_{l=1}^{L} \alpha_l d_l^{(k)}(\bar{v}_l, v_l^{(k)}), \tag{5}$$

where, $d_1^{(k)}$is the distance between the blob's low-level-feature descriptors and the centroids and α_l the elements of the set of weighting coefficients to optimise.

3. Multi-objective Optimisation and *Pareto Optimum*. The challenge consists of optimising the set of formulated objective functions and therefore, optimising α_l, in order to represent every semantic object with a suitable mixture of low-level-feature descriptors. However, two aspects need to be taken into consideration: (i) single optimisation of each *object function* may lead to biased results; (ii) the contradictory nature of low-level-feature descriptors should be considered in the optimisation process. The existence of several *objective functions* ensures better discrimination power compared to using a single *objective function*. Consequently, a set of compromised solutions, known as *Pareto-optimal solutions* are generated using the multi-objective optimisation-strategy that relies on a local search algorithm. Individual *Pareto-optimal solutions* cannot be consider better than the others without further consideration. Therefore, a set of conditions are allocated to choose the most suitable *Pareto-optimal solution*, (i) to minimise the *object functions* of the negative training samples, (ii) to maximise the *object functions* of the positive training samples and (iii) the sum of the elements of A must fulfil $\sum_{l=1}^{K}\alpha_l = 1$.

Once the requirements have been set, a *decision making* step must take place, to find a unique solution which minimise the ratio between *(i)* and *(ii)*:

$$\min \frac{\sum_{k=1}^{K} D_{+}^{(k)}(V^{(k)}, \bar{V}, A_s)}{\sum_{k=1}^{K} D_{-}^{(k)}(V^{(k)}, \bar{V}, A_s)}, s = 1, 2, ..., S \qquad (6)$$

where $D_{-}^{(k)}$ and $D_{+}^{(k)}$ are the distances over positive and negative training samples respectively, while, A_s is the s^{th} in the set of *Pareto-optimal solutions*, and S is the number of available *Pareto-optimal solutions*.

4. Similarity Matching Function. The optimised *Multi-feature* matching function for any blob example is calculated by:

$$D_{MOO}(V, \bar{V}, A) = \sum_{l=1}^{L} \alpha_l d_l(v_l, \bar{v_l}), \qquad (7)$$

the resulting values $D_{MOO}(V, \bar{V}, A)$ represent the likelihood of a blob to contain a certain concept, in our case *Person* or *Car* (concepts considered positive while computing Eq. (6)).

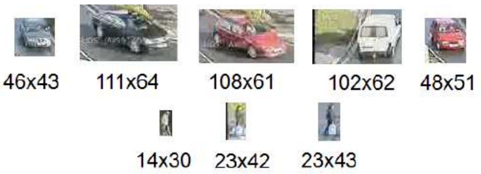

Figure 3: Representative set of blobs from the *Ground truth*, which resolution is also presented.

5 EXPERIMENTAL EVALUATION

AVSS 2007 dataset [2] was used to evaluate the presented surveillance video retrieval approach providing indoor and outdoor videos summing a total of 35000 frames. For evaluation purposes, three outdoor videos were selected at different levels of difficulty. A total of 1377 objects were included and manually annotated in the ground truth, of which 50% of objects were annotated as “Cars” against 10% annotated as “Person” and the remaining 40% were annotated as “Unknown”. Instead of ignoring the blobs labelled as “Unknown”, our dataset included these blobs to explicitly study the effect of noise on the performance of the retrieval models. An overview of the dataset used for the evaluation of the proposed framework is presented in Fig. 3. Besides less than 6% of the ground truth was selected to form the training dataset which was used to train the *Multi-objective optimisation* component.

5.1 MPEG-7 Visual Feature Extraction

In this section, a short description of the set of selected MPEG-7 descriptors, chosen by their robustness, compact representation and significance for human perception is presented.

Colour Layout Descriptor (CLD) is a very compact and resolution-invariant representation of the spatial distribution of colour in an arbitrarily-shaped region (Sikora, 2002).

Colour Structure Descriptor (CSD) describes spatial distribution of colour in an image, but unlike colour histograms, *CSD* also describes local colour spatial distribution.

Dominant Colour Descriptor (DCD) describes global as well as local spatial colour distribution in images for fast search and retrieval. *DCD* provides a description on the distribution of the colour within an analysed image by storing only the a small number of representative colours or *dominant colours*.

[2] http://www.eecs.qmul.ac.uk/ãndrea/avss2007_d.html

Edge Histogram Descriptor (EHD) provides a description for non-homogeneous texture images and captures the spatial distribution of edges whilst providing ease of extraction, scale invariance and support for rotation-sensitive and rotation-invariant matching.

5.2 Experimental Evaluation of Particle Swarm Optimisation

The PSO model implemented is a combination of cognitive and social behaviour. The structure of the PSO is fully connected in which a change in a particle affects the velocity and position of other particles in the group as opposed to partial connectivity, where a change in a particle affects the limited number of neighbourhood in the group. Each dimension of the feature set is optimized with 50 particles. The size of the SOM network is pre-fixed with the maximum number of training samples to be used in the network. The stopping criteria threshold is experimentally determined for different individual feature space. The value of the threshold indicated the closeness in solving the optimization problem. In Fig. 4, performance comparison of PSO based retrieval is evaluated against different kernels of Support Vector Machines (SVM). As it is noted from the results, the performance of the classifier varies according to the feature space. This could be largely attributed to the extraction of different features and the matching functions involved in these distinct spaces. From the results, we can see that CLD space is quite optimal for retrieving the concept ’Car’, while for concept ’Person’, the retrieval performance drops beyond recall at 0.5. Similar interpretations could be extrapolated from CSD feature space where the performance for retrieving ’Car’ is higher than for concept ’Person’.

5.3 Evaluation of Optimal Combination of Low-level Feature Space

In Fig. 5, precision-recall curve for the concept ’Car’ is presented with performance comparison of PSO algorithm with optimal and primitive low-level feature fusion technique. As it can been, the primitive combination of the feature vector, drops in retrieval performance at lower recall, but remains competitive over mid-range recall values. On the other hand, the optimal combination achieves improved retrieval performance for lower recall values. However, the retrieval performance drops over mid-range recall and for 0.83 recall both techniques achieve same precision. Interestingly, from the study of results for the concept “Person”, it can be easily noted that, the performance of optimal combination of feature vector is much better compared to its primitive counter part, refer to Fig. 6. The improved performance of the optimal combination of low-level features could be attributed to the fact that, the optimisation technique determines appropriate weights for all concepts in the multi-descriptor space, achieving a overall balanced solution. With the aim of obtaining global perfor-

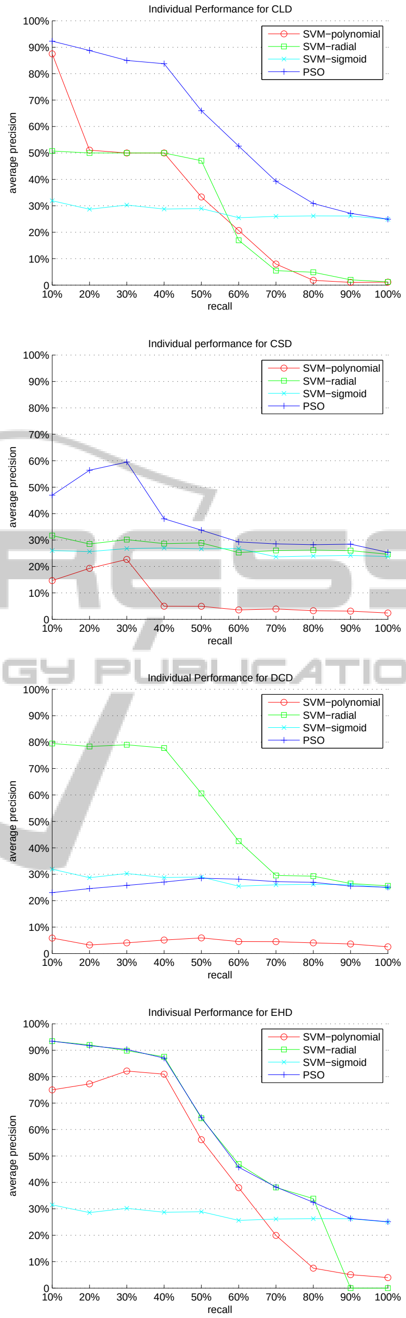

Figure 4: Performance of Particle Swarm and Kernal Machines in Individual Feature Space for Concepts ’Car’ and ’Person’.

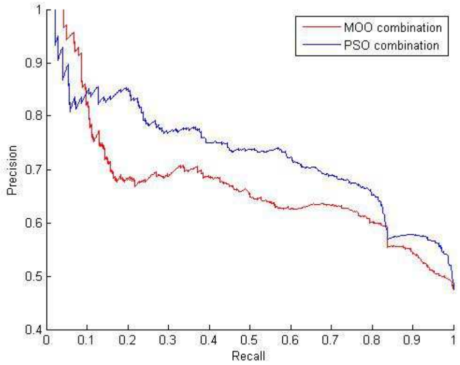

Figure 5: Precision-Recall curve for Surveillance Object Retrieval using optimal combination and primitive combination for concept *Car*.

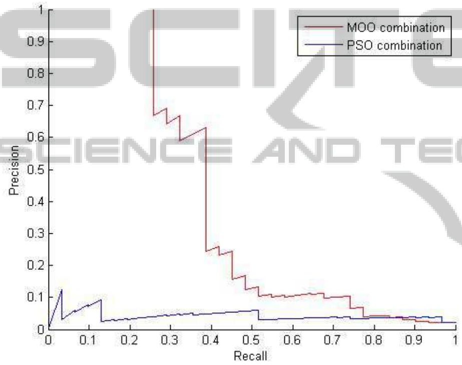

Figure 6: Precision-Recall curve for Surveillance Object Retrieval using optimal combination and primitive combination for concept *Person*.

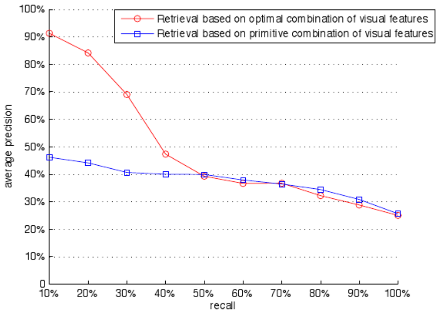

Figure 7: Average Precision-Recall curve for Surveillance Object Retrieval using optimal combination and primitive combination across both concepts.

mance, we clearly note that the optimal combination of low-level feature space performs better compared to the primitive combination as highlighted in FIg. 7. The average performance obtained over two concepts for optimal combination is nearly 40% more than the primitive combination at 0 recall. However, from 50% recall both technique provide similar results with respect to average precision-recall.

6 CONCLUSIONS & FUTURE WORK

In this paper, an optimal combination of low-level descriptor space was presented for surveillance object retrieval. The optimal combination of multi-descriptor space was evaluated against individual descriptor space using Particle Swarm Classifier and three support vector machines kernels. In addition, an evaluation of optimally combined feature space was evaluated against a primitive combination of feature space. Moreover, a detailed study of the results was carried out. As noted in the results, the proposed retrieval framework achieves 40% improvement over the primitive combination and more importantly consistent performance is obtained across different concepts. For the future work we will extend the study to include more concepts and novel non-MPEG-7 visual features. Similarly, a relevance feedback module will be included for online training of the system.

REFERENCES

Chandramouli, K. and Izquierdo, E. (2010). *Image Retrieval using Particle Swarm Optimization*, pages 297–320. CRC Press.

Eberhart, R. and Shi, Y. (2001). Tracking and optimizing dynamic systems with particle swarms. *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on*, 1.

Kennedy, J. and Eberhart, R. C. (2001). *Swarm intelligence*. Morgan Kaufmann.

Mojsilovic, A. (2005). A computational model for color naming and describing color composition of images. *Image Processing, IEEE Trans.*, 14(5):690–699.

Sikora, T. (2002). The MPEG-7 Visual standard for content description-an overview. *Circuits and Systems for Video Technology, IEEE Trans.*, 11(6):696–702.

Soysal, M. and Alatan, A. (2003). Combining MPEG-7 based visual experts for reaching semantics. *Visual Content Processing and Representation*, pages 66–75.

Stauffer, C. and Grimson, W. (2000). Learning patterns of activity using real-time tracking. *Pattern Analysis and Machine Intelligence, IEEE Trans.*, 22(8):747–757.

Xu, R. and II, D. W. (2005). Survey of clustering algorithms. *IEEE Trans. Neural Network*, 6(3):645–678.

Zhang, Q. and Izquierdo, E. (2007). Combining low-level features for semantic inference in image retrieval. *Journal on Advances in Signal Processing*.