# MULTI-FEATURE FUSION FOR SURVEILLANCE VIDEO INDEXING*

*Virginia Fernandez Arguedas, Qianni Zhang, Krishna Chandramouli, 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, UK.
{virginia.fernandez, qianni.zhang, krishna.chandramouli, ebroul.izquierdo}@eecs.qmul.ac.uk

## ABSTRACT

In this paper, we present a part of surveillance centric indexing framework aimed at studying the performance of multi-feature fusion technique for indexing objects from surveillance videos. The multi-feature fusion algorithm determines an optimal metric for fusing low-level descriptors extracted from different feature space. These low-level descriptors exhibit a non-linear behaviour and typically consist of different similarity metrics. The framework also includes a motion analysis component for the extraction of objects as blobs from individual frames. The proposed framework, in particular the multi-feature fusion algorithm is evaluated against kernel machines for indexing objects such as car and person on AVSS 2007 surveillance dataset.

***Index Terms*—** multi-objective optimisation, multidescriptor space, pareto optimisation, object indexing, support vector machines, motion analysis, MPEG-7 features, surveillance application.

## 1. INTRODUCTION

Recent technological developed coupled together with people's concern for safety and security have caused a wide spread application of Closed Circuit Television (CCTV) camera which has been installed in large-scale for surveillance monitoring. This large-scale deployment of CCTV generates a huge volume of video footage as the cameras operate 24/7. With such an exponential increase in video information, there exists critical need for the development of automatic and intelligent indexing schemes 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 [1]. Many of these techniques generate satisfactory results for general datasets such as movies, sports and news. However, the challenge of indexing surveillance objects remains a largely an open issue.

*THE RESEARCH WAS PARTIALLY SUPPORTED BY THE EUROPEAN COMMISSION UNDER CONTRACT FP7-216444 PETAMEDIA.

Among the proposed approaches in the literature, visual appearance based indexing has gained much popularity. To this end, the range of visual features has been used to index surveillance objects include, colour histograms from different colour space, Gabor filters, MPEG-7 based colour, texture and shape descriptors. In many of these approaches authors consider a single low-level descriptor to provide a high-level degree of distinguishability among objects. Even when considering multiple low-level descriptors, authors often neglect the non-linearity of the descriptor space and combine these features in a linear manner. The need for multi-feature descriptors is motivated by the attempt of generating more robust and complex representation. To this end, a large number of different features are used to represent objects obtained from surveillance videos. However, in doing so its critical to consider different feature characteristics [2]. The combination of low-level-features to obtain higher order representations have been addressed over the years in pattern recognition. For instance, in [3, 4] authors proposed approaches that used combination of multiple low-level features to index images. However, to the best of our knowledge, such feature fusion approaches hasn't been applied for indexing objects from surveillance video datasets.

In this paper, we present a part of surveillance centric indexing framework aimed at studying the performance of multi-feature fusion technique for indexing objects from surveillance videos. The indexing schema uses different low-level features, which exhibit a non-linear behaviour and typically consist of different similarity metrics. The multi-feature fusion algorithm presented, determines an optimal metric for fusing these visual descriptors extracted in different feature space. The framework also includes a motion analysis component for the extraction of objects as blobs from individual frames. The framework is evaluated against conventional kernel machines for performance comparison.

The rest of the paper is organised as follows. The proposed object indexing framework is presented in Section 2, followed by a brief discussion on the Motion Analysis component in Section 3. The multi-objective optimisation tech-

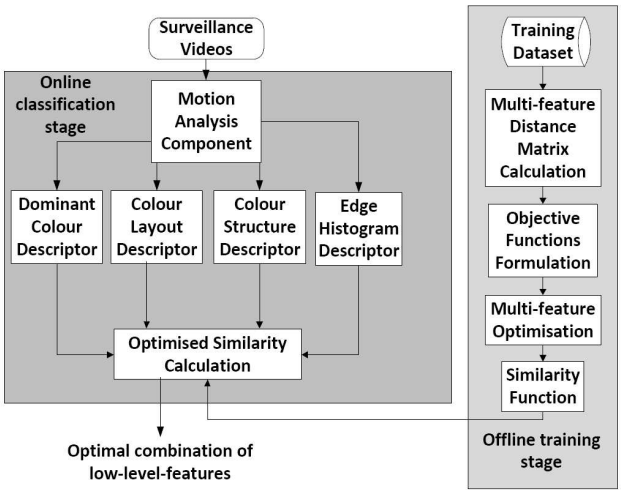


**Fig. 1**. Framework overview

nique is presented in Section 4. Experimental results obtained on applying the proposed approach over a surveillance video dataset is presented in Section 5. Finally, conclusions and future work are briefly discussed in Section 6.

## 2. SURVEILLANCE CENTRIC INDEXING FRAMEWORK

The proposed surveillance centric indexing framework is presented in Fig.1. The framework consists of two stages namely the online classification and offline training mode. The offline training stage consists of the multi-feature fusion algorithm, which is used to create visual models to enable indexing. In the online classification phase, the video is subjected to motion analysis component to extract the blobs from the videos. The motion analysis component is based on Stauffer and Grimson approach [5]. A more detailed description of the motion analysis component is presented in Section 3. Similarly, the multi-feature fusion algorithm is discussed in detail in Section 4. The multi-feature fusion algorithm calculates distance metric according to the feature space for each of the object blob extracted. Following which a multi-objective optimisation technique is applied to obtain an optimal mixture of the previously calculated low-level features that represents a certain object in the best possible manner.

## 3. MOTION ANALYSIS COMPONENT

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. A three-step real-time *Motion Analysis Component* is presented

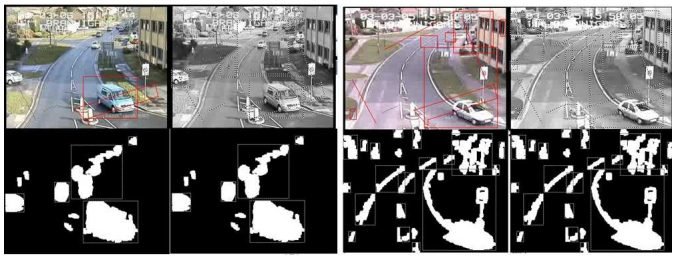

**Fig. 2**. Motion analysis component results. Background subtraction and spatial segmentation techniques results can be observed for two different problematic situations as low quality image (left) and videos with camera movement (right)

to procure individual blobs to the *Feature Extraction Component*.

First, an adaptive background subtraction technique based on Stauffer and Grimson algorithm [5] is performed to remove all the redundant information of the surveillance videos, allowing a faster analysis and providing robustness against external factors, such as changes in illumination or camouflage. Adaptive background subtraction algorithm is a two-step process (i) modelling a background as a mixture of Gaussians and (ii) modelling each pixel of an image as a weighted mixture of Gaussians and classifying it into foreground or background according to the persistence and variance of each of the Gaussians of the mixture. Thus, pixels are classified as foreground if their values do not fit the background distributions formerly calculated. Second, object spatial segmentation is performed grouping the resulting Gaussian mixtures. Consequently, a two-pass connected component algorithm assuming an 8-connection is applied. As a result, foreground moving objects are isolated. Third, temporal segmentation is performed establishing the correspondence of the spatially segmented objects between frames using a linearly predictive multiple hypothesis tracking algorithm based on a set of Kalman filters. Moreover, Kalman filters are used to predict the tracks related to each frame as well as the assignment between the available tracks and the detected blobs in each frame. Despite many advantages of the use of motion analysis component, as highlighted in Figure 2, object detection from surveillance video is affected by a lot of noise generated from (i) the low quality of the image; (ii) lack of contrast or the image blurring due to the camera motion.

## 4. MULTI-OBJECTIVE OPTIMISATION TECHNIQUE

Since single low-level feature descriptors are not capable of interpreting human understanding, a joined combination of different low level feature descriptors is provided. However, their different nature, different metrics and non-linear behaviours make their combination difficult. The challenge in *Multi-*

*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 [3]. First, a distance matrix between each blob and feature is calculated. Second, a global multi-feature weighted metric is formulated as objective function for each training blob. Third, the contradictory nature of the low-level feature descriptors may display different interests in objective functions. To obtain a balanced and compromised general solution which considers all the conditions, *Pareto-optimal solutions* are calculated from the set of objective functions of the training blobs. Fourth, a unique solution is calculated applying several constraints.

**Distance matrix calculation**: Four MPEG-7 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 by the *Colour Layout Descriptor*, *Scalable Colour Descriptor*, *Dominant Colour Descriptor* and *Edge Histogram Descriptor*, a centroid is calculated for each of the low-level-feature descriptors generating a virtual centroid vector called $\bar{V} = (\bar{v}_{CLD}, \bar{v}_{SCD}, \bar{v}_{DCD}, \bar{v}_{EHD})$. 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.

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

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

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

**Multi-objective optimisation and Pareto optimum**: The challenge consists of optimising the set of formulated objective functions and therefore, optimising $\alpha_l$, 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, *(a)*; (ii) to maximise the *object functions* of the positive training samples, *(b)*; 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 *(a)* and *(b)*:

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

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*.

**Similarity matching function**: The optimised *Multi-feature* matching function for any blob example is calculated using Equation 3, where 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*.

$$D_{MOO}(V, \bar{V}, A) = \sum_{l=1}^{L} \alpha_l d_l(v_l, \bar{v}_l), \qquad (3)$$

## 5. EXPERIMENTAL RESULTS

AVSS 2007 dataset [1] was used to evaluate the presented surveillance video indexing approach providing indoor and outdoor videos summing a total of 35000 images. For evaluation purposes, three outdoor videos, with a total of 13400 images, were analysed with variable lighting conditions as well as different levels of difficulty. The surveillance footage includes several challenges such as noise, low quality image, camera movement or blurring increasing the difficulty of its analysis.

### 5.1. Quantitative Performance Evaluation

To investigate the performance of our surveillance video indexing approach a ground truth was developed selecting a relatively small sized set of blobs extracted from the dataset and manually annotated with two predefined concepts, *Car* and *Person* (see Figure 3). A total of 1376 objects were included and annotated in the ground truth. Besides, the ground truth was partially selected to form the training dataset which was used to train the *Multi-objective optimisation* component. Less than a 6% of the ground truth was selected for the training dataset, where 90% of the objects were annotated as *Car* against the 10% as *Person*.

Formerly, all the objects were spatially and temporally segmented by the *Motion Analysis Component* from surveillance videos. In order to evaluate the performance of the

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

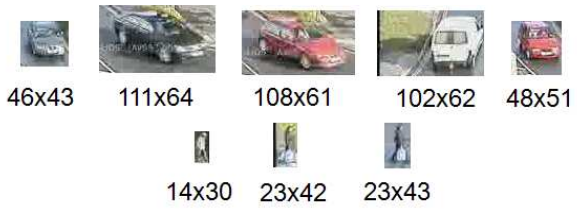


**Fig. 3**. Representative set of blobs from the *Ground truth*, which resolution is also presented

*MOO*, four MPEG-7 features with different feature spaces were extracted: *Colour Layout Descriptor (CLD)*, *Scalable Colour Descriptor (SCD)*, *Dominant Colour Descriptor (DCD)* and *Edge Histogram Descriptor (EHD)* [6]. All these features were chosen by their robustness, compact representation and significance for human perception. In addition, we quantitatively evaluated the results obtained indexing the extracted moving objects considering all its low-level-features equally important and using SVMs to classify them. Furthermore, the improvement provided by *Multi-objective Optimisation Technique (MOO)* as an optimal linear combination of the low-level features to index the surveillance objects was studied over the dataset.

The selected *MPEG-7 features* were computed to index the extracted objects from the surveillance videos giving all the features the same relevance. In order to study their efficiency, *Support Vector Machines (SVM)* were applied [2]. First, a model for Car concept was created using the training dataset. Second, the distances of the Car model were computed. In this calculation, all the MPEG-7 features were considered equally relevant and their different feature spaces were not taken into account. The obtained results are shown in Table 1.

*Multi-objective Optimisation Technique* was applied in order to provide an optimal linear combination for the low-level-feature descriptors while considering that each feature have a different feature space. In order to study its efficiency, a retrieval process was applied using the optimal low-level-feature descriptor as an index. The obtained results are shown in Table 1.

Results provided by *SVM* reveal a considerable F-measure for the concept Car, however, its performance for the concept Person is insufficient. A reason for Person results can be related to the sparseness of the concept within the ground truth, where the concept Person covers a 3% of the total. *MOO* was applied to consider the different feature spaces of the extracted low-level features. Its results show a reasonable improvement for Car and Person concept. Even thought the sparseness of the Person concept in the surveillance ground truth is still an open issue.

[2]The module used to compute the *Support Vector Machines* is based on Cornell University's module, www.cs.cornell.edu/People/tj/svm_light

| Concepts | F-measure (%) | | | | | |
|---|---|---|---|---|---|---|
| | CLD | EHD | SCD | DCD | SVM | MOO |
| PERSON | 4.47 | 19.14 | 7.76 | 63.82 | 7.92 | 25.35 |
| CAR | 5.92 | 68.48 | 65.38 | 7.27 | 45.69 | 64.43 |

**Table 1**. Performance comparison of MPEG-7 features with SVM and MOO fusion techniques

## 6. CONCLUSIONS & FUTURE WORK

In this paper, a multi-feature fusion algorithm was presented for indexing objects from surveillance videos. MPEG-7 visual features were applied to obtain an optimal combination of the feature metrics. The performance evaluation study conducted against support vector machines indicate a 20% improvement in indexing performance. The future work will focus on extending the multi-descirptor feature space beyond MPEG-7 and also will focus on the use of local features to improve the object indexing schema. The surveillance centric framework will further be extended to include high-level feature space such as motion velocity, motion acceleration and motion correspondence to improve the indexing process.

## 7. REFERENCES

[1] Chandramouli, K., Izquierdo, E.: Image Retrieval Using Particle Swarm Optimisation. In: Book Advances in Semantic Media Adaptation and Personalisation, vol. 2, pp. 297-320 (2010).

[2] Mojsilovic, A.: A computational model for color naming and describing color composition of images. IEEE Transactions on Image Processing, vol.14, no.5, pp.690-699 (2005).

[3] Zhang, Q., Izquierdo, E.: Combining low-level features for semantic inference in image retrieval. In: EURASIP Journal on Advances in Signal Processing (2007).

[4] Soysal, M., Alatan, A.A.: Combining MPEG-7 basedvisual experts for reaching semantics. In: 8th International Workshop on Visual Content Processing and Representation, vol.2849, pp. 66-75. Madrid, Spain (2003).

[5] Stauffer, C., Grimson, L.: Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.22, pp. 747-757 (2000).

[6] Sikora,T.: The MPEG-7 visual standard for content descriptor - An overview. In: IEEE Transactions on circuits and systems for video technology, vol.11, no.6, pp.696–702 (2002).