I.J. Image, Graphics and Signal Processing, **2015, 8, 1-12**
Published Online July 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2015.08.01

A Novel Object Position Coding for Multi-Object Tracking using Sparse Representation

Mohamed ELBAHRI[1,*]
[1]Department of Computer Science, Djillali Liabes University, Sidi Bel-Abbes, Algeria
E-mail addresses: elbahri82_m@yahoo.fr

Kidiyo KPALMA[2], **Nasreddine TALEB**[3], **and Miloud CHIKR EL-MEZOUAR**[3]
[2]UEB INSA IETR Dépatement Image et Automatique, 35708 Rennes, France
[3]Department of Electronics, Djillali Liabes University, Sidi Bel-Abbes, Algeria
E-mail addresses: Kidiyo.Kpalma@insa-rennes.fr, ne_taleb@univ-sba.dz, chikrelmezouar@univ-sba.dz.

Abstract—Multi-object tracking is a challenging task, especially when the persistence of the identity of objects is required. In this paper, we propose an approach based on the detection and the recognition. To detect the moving objects, a background subtraction is employed. To solve the recognition problem, a classification system based on sparse representation is used. With an online dictionary learning, each detected object is classified according to the obtained sparse solution. Each column of the used dictionary contains a descriptor representing an object. Our main contribution is the representation of the moving object with a descriptor derived from a novel representation of its 2-D position and a histogram-based feature, improved by using the silhouette of this object. Experimental results show that the approach proposed for describing moving objects, combined with the classification system based on sparse representation provides a robust multi-object tracker in videos involving occlusions and illumination changes.

Index Terms—Multi-object tracking, Object representation, Orthogonal matching pursuit, Sparse representation, Classification.

I. Introduction

Video surveillance systems are widely used in public places. To make it more effective, the technical parts of computer vision emerging as a tracking of moving objects permit now to recognize, index, and track the path of everything moving in a video. In tracking systems, the challenge is to ensure the persistence of the object's identity by addressing the issues perturbing tracking algorithm such as target position changes, occlusion and illumination changes.

In the present study, tracking is formulated as a classification system. An online dictionary composed of sub-classes is used: each element of a sub-class representing the same object. Each column of the dictionary contains a descriptor of a detected object in previous frames. Initially, the dictionary is empty; it is then updated during the detection of any moving object. A new sub-class is created in the dictionary to label any object appearing for the first time, and the system classifies the object into the same sub-class when it reappears again in the next frames. Sparse representation (SR) is used in different areas of computer vision and image processing [1,2,3,4,5,6]. The main goal of using sparsity is to represent [7] or classify an input signal by choosing a subset of models or features from data acquired during learning rather than the data itself [8,9]. In our work, we use sparse representation to classify the input test object. Object recognition can be expressed as a classification system through a linear regression model. This problem of SR-based classification can be addressed by computing the ℓ^1 minimization [8,10,11,12]. In the present study, Orthogonal Matching Pursuit (OMP) is used to classify objects according to the sparse solution. This algorithm, proposed by Y. Pati [13], is an extension of the matching pursuit algorithm developed by Mallat et Zhang [14]. This algorithm aims at determining the sparse coefficients. The first step of OMP, is to select the column of the dictionary which generates the largest inner product with the current feature vector. Hence, the largest inner product generated by this vector with the i^{th} column of the dictionary means that this column represents the same object in one of the previous frames.

The contribution of this work consists of the representation of objects. To get the largest product between two vectors representing respectively, two objects having similar colors and sharing the same location, we propose to represent moving objects, with two major parts. The first part represents the appearance of the object and the second its position. In our experiments, we focus on the people tracking video. Hence the object body is divided into two parts: the upper part and the lower one [15]. The motivation is that in everyday's life, people wear upper part clothes of colors different than those of lower part clothes. Moreover, instead of using bounding boxes like in the existing approaches, this description is then improved by using masks derived from the silhouette of the object to get the appearance feature, and by using a vector to represent the object position.

In the rest of this paper, we discuss the following issues: Section II presents some related work. In section III, we detail the proposed scheme to build the object descriptor. In section IV we give a brief introduction to sparse representation and summarize the OMP algorithm. Section V presents and discusses the experimental results while section VI concludes the paper.

II. Related work

In the literature, several approaches have been developed for object tracking [16,17,18,19]. In these methods, some authors locate the target object using information from the future frames and process for detection over large temporal windows. Black et al. [20] based their work on the optical flow to represent target objects. Jepson et al. [21] used a Gaussian mixture model of pixels to represent objects via an online expectation maximization (EM) algorithm. David et al. [21], implemented tracking with a particle filter instead of blobs of pixels. Avidan [23] trains a discriminative model using a Support Vector Machine (SVM) to detect moving objects and separate them from the background. In [24], some methods are proposed using classifiers to distinguish the moving objects.

Most recent research works have focused on sparse representation to deal with the problem of object tracking [25,26,27,28,29]. Lu et al. [15] present some experimentation of OMP and LARS [30] for tracking. Furthermore, the authors in [31] propose a kernel-based method with multikernel fusion into SR for robust visual tracking within the particle filter framework.

The proposed approach consists of two major steps: detection and recognition of moving objects. In this context, the adopted strategy is to describe moving objects by their appearance features and by their positions in the scene. For more accuracy, all of the information belonging to the background is removed. So after segmentation, the color histogram is computed and used as the first part of the descriptor. In section III-B, experimental results are given to illustrate how much background information can disturb the appearance description. To solve this problem, the proposed solution computes the histograms from a mask derived from the silhouette of the object.

The object descriptor, containing the appearance feature only, cannot assure the persistence of the objects identity. Integrating the object position in the descriptor will indeed contribute to improve the recognition. In several works [15,32,33], the position is encoded using the simple coordinates of the object for its representation; others use position to compute the distance between objects and the camera. In the present work, the position is encoded as a vector derived from the whole frame; this vector is concatenated with the vector representing the appearance feature.

III. Object detection and representation

Within the object-tracking framework, the detection and representation of the moving object are crucial steps. The objectives of these steps consists of detecting as accurately as possible moving object, and to represent it, with a model robust to illumination changes and to the occlusion problem.

A. Object Detection

In the case of a fixed camera, using background subtraction [34] is an effective way for motion detection. The result of this operation may however contain some undesirable noise due to background changes as illumination changes and undesirable movements that must be filtered. Other approaches are proposed in the literature such as "Mixture of Gaussian" to separate the foreground from the background [35], or some classification systems using features such as in Hog [36] and in Haar [37,38].

Fig.1 illustrates a background subtraction by computing the absolute difference between the background (Fig.1.a) and another frame acquired at a particular time (Fig. .b). After thresholding the result of the background subtraction, the bodies shapes are visible in Fig.1.(c).

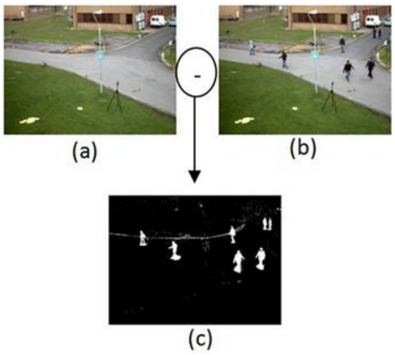

Fig. 1. (a) Background reference. (b) Observed scene. (c) Thresholded result of the background subtraction.

It is to note that in Fig. 1.c, the result of the subtraction does not include just the bodies' shapes but some noise also. To improve the extraction of the detected moving objects bodies, mathematical morphology operations are used [39,40]. Erosion is applied to eliminate noise such as the one characterized by the white line due to the movement of the horizontal ribbon. This process uses a structuring element that consists of a (7x7) matrix of 0's and 1's values. The proposed structuring element eliminates the scattering pixels or those constituting horizontal shapes. However, erosion generates an adverse effect and dissociates bodies' limbs as can be seen in Fig. 2 .(b), so we apply a dilation with another structuring element consisting of a (9x9) matrix to reconstruct the bodies shapes.

It was observed that the background subtraction technique is not very robust for the detection of the bodies shapes, when the scene is very crowded and the objects overlap. An approach has been proposed in the section III. D which aims to deal with the overlapping objects.

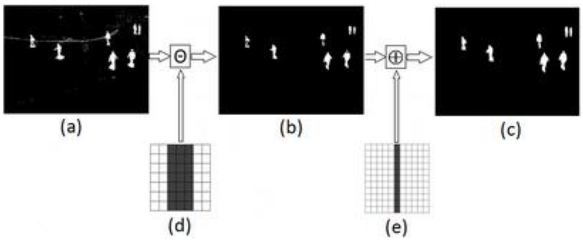

Fig. 2. (a) Background subtraction result. (b) Erosion result. (c) Dilation result. (d) Structuring element used for erosion. (e) Structuring element used for dilation.

B. *Object Description*

After the detection of the object's shape, edge detection is applied to the binary image resulting from the background subtraction to delineate the silhouette of the body of the moving object. In the proposed approach, the histogram is derived from this silhouette, while the majority of techniques used in this field calculate the histogram from bounding box of the object. So the proposed histogram is more accurate and more representative of the object. Fig. 3 represents the obtained edge detection.

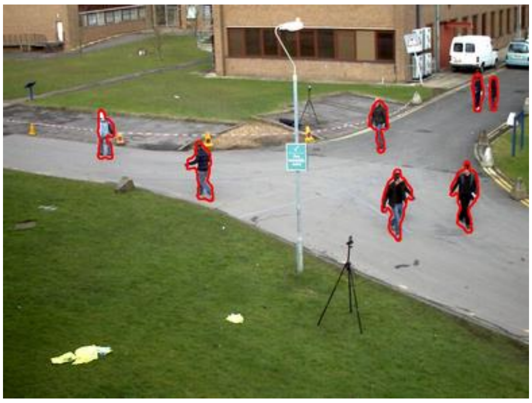

Fig. 3. Detected edges superposed on the original image.

For each detected object, a rectangular mask is extracted from the binary image (Fig.2.c) . Fig. 4 shows a zoom of a detected object (Fig.4.a) as well as the mask extracted for the same object (Fig.4.b). A 0 value represents the background (black), while a 1 value represents pixels of the object body silhouette (white).

Using silhouettes as masks of the objects makes histograms more accurate since only pixels representing the moving objects are accounted. In the following example (Fig.5), the histograms are computed from two bounding boxes representing the same person in different frames. From each bounding box, the histogram is computed in two ways: the first using the mask and the second without it.

After that, the similarity is measured between each pair of histograms. Figures 5.a and 5.c show the same person Black pixels represent 42% of the entire box, hence, the background pixels distort the curve of the graph representing the histogram.in different frames and in locations with different backgrounds. Fig.5.b shows only a person without the background pixels.

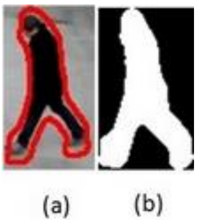

Fig. 4. (a) Contour of the object (b) Corresponding mask of the object.

It is noted that in the graph corresponding to Fig.5.a, the peak at level L_a represents the background pixels and it is higher than the one at level L_b of the graph corresponding to the Fig.5.b. The peak at level L_a is reduced when the histogram is computed using the mask. The goal of this approach is to improve the similarity between two histograms representing the same person and subsequently reducing the effect of the background. As can be seen, for example, the green graphs in figures 5.b and 5.d are more similar than those of figures 5.a and 5.c. This is confirmed by the similarity measured in both situations.

The degree of similarity between two images can be measured with the correlation between their histograms. Let $h_1 and\ h_2$ be the histograms of two images I_1 and I_2, respectively. The correlation between these histograms is defined as:

$$d_c(h_1^c, h_2^c) = \frac{\sum_{i=1}^{N} h_1^c(i) h_2^c(i)}{\sqrt{\sum_{i=1}^{N} h_1^c(i)^2 \sum_{i=1}^{N} h_2^c(i)^2}} \quad (1)$$

where $c \in \{R, G, B\}$ with $0 \leq d_c(h_1^c, h_2^c) \leq 1$, and if $d_c(h_1^c, h_2^c) \approx 1$ then h_1 and h_2 are similar.

Let $h_1^c, \hat{h}_1^c, h_2, \hat{h}_2^c$, be the histograms corresponding to Fig.5.a, Fig.5.b, Fig.5.c and Fig.5.d, respectively. The average correlation between the histograms of the three color channels of the example in Fig.5 is as follows: $\overline{d_c(h_1^c, h_2^c)} = 0{,}53\ and\ \overline{d_c\left(\hat{h}_1^c, \hat{h}_2^c\right)} = 0{,}83.$ It is to notice that $\overline{d_c\left(\hat{h}_1^c, \hat{h}_2^c\right)} > \overline{d_c(h_1^c, h_2^c)}$.

The correlation of histograms representing the same person in different scenes with different backgrounds using silhouette masks is greater than that using a bounding box. Based on this observation, we confirm that using the silhouette as a mask in the histogram construction is highly beneficial.

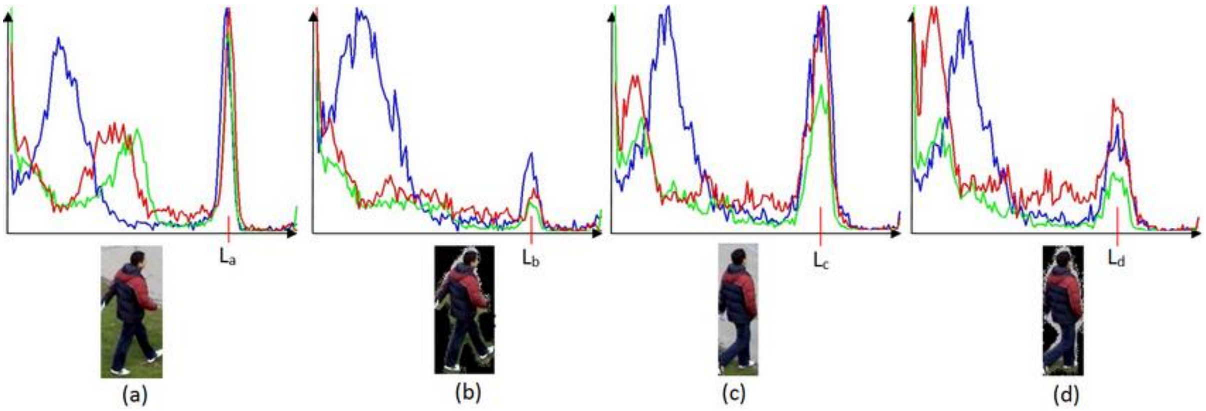

Fig. 5. Example of Color histograms of the images.

C. *Position Feature*

The position of the moving objects contributes to the objects recognition. It is represented in literature with the coordinates of the center of the object. The classification algorithm used in this work for recognition is based on the inner product for the selection of the most correlated column in the dictionary with the input test object. However, the inner product of simple coordinates cannot provide information about the distance between two objects: two objects can be highly correlated but far from each other and close objects can be weakly correlated.

To illustrate this statement, let $P_i\ s$ be the positions of three objects at t_1, and $\hat{P}_j\ s$ be the positions of the same three objects at t_2 , respectively, where $P_i, \hat{P}_j \in \mathbb{R}^{2\times1}$ and $i, j \in \{1,2,3\}$. Following the example in Fig.6, we obtain the following equation:

$$\forall\ i, argmax_j \langle P_i, \hat{P}_j \rangle = 3 \tag{2}$$

where $\langle P_i, \hat{P}_j \rangle = \binom{x_i}{y_j} \times \binom{\hat{x}_i}{\hat{y}_j}^T$. After observing the results of the inner product of the different positions in Table1, we confirm formula (2), and the largest inner product is the one generated with the position $\hat{P}_3$ corresponding to the farthest object from the origin.
However, we need to represent the position such that:

$$\forall\ i, argmax_j\ \langle P_i, \hat{\hat{P}}_j \rangle = i \tag{3}$$

To satisfy the formula (3), a novel representation of the position is proposed in this work. The idea is to create a structure to encode the position of each detected object. This structure is derived from a matrix $Q \in\ \mathbb{R}^{m,n}$ created for each detected object. This matrix, called the position matrix, is used to sample the image in small areas.

Fig.7 shows the position matrices of three detected objects. The frame in Fig.7.a.1 is acquired at t_1 and the position matrix in Fig.7.a.2 is used to represent the position of the person that we call "person1". figures 7.b.1 and 7.c.1 show the frame acquired at t_2, and the position matrices corresponding to two persons ("person1" and another that we call "person2") are illustrated by figures 7.b.2 and 7.c.2, respectively. It is noted that in each position matrix, the red cell corresponds to the object center. For the other cells, their color is modified according to the distance between them and the red one.

To implement the used approach, matrix $Q \in\ \mathbb{R}^{m,n}$ initially contains only zeros. We set the value $\alpha \in \mathbb{R}^+$ to cell $Q_{i,j}$ corresponding to the coordinates $(x\ , y)$ of the object center where $= \left\lfloor \frac{x}{\frac{h}{m}} \right\rfloor, j = \left\lfloor \frac{y}{\frac{w}{n}} \right\rfloor$, with $i, j\ \in \mathbb{N}$, h is the image height and w its width.

The other cells values are:

$$Q_{i',j'} = \begin{cases} \alpha - \gamma\vartheta\ if \geq 0 \\ \quad else\ 0 \end{cases} \tag{4}$$

where $\gamma = \max(|i' - i|, |j' - j|)\ and\ \vartheta \in \mathbb{R}: 0 < \vartheta < \alpha$

The position matrix must be modified in order to be concatenated with the vector encoding the histogram representing the appearance feature. A position vector is created by applying a zig-zag scan on the position matrix. Figures 8.a, 8.b and 8.c show the position vectors corresponding to the position matrices of figures 7.a.2, 7.b.2 and 7.c.2, respectively.

Table 1: Result of the inner products of the objects' positions according to Fig. 6

$\langle P_i, \hat{P}_j \rangle$	j=1: (30,15)	j=2: (60,30)	j=3: (90,50)
i=1: (20,15)	825	1650	**2250**
i=2: (60,25)	2175	4350	**6650**
i=3:(100,45)	3675	7350	**11250**

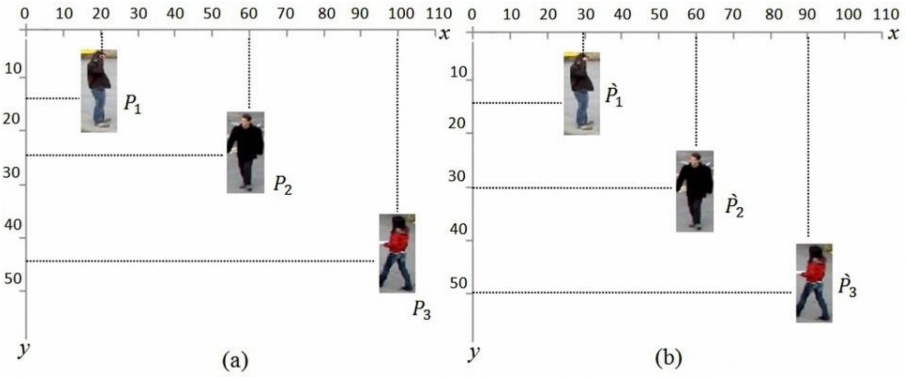

Fig. 6. (a) Objects positions at t_1. (b) Objects positions at t_2..

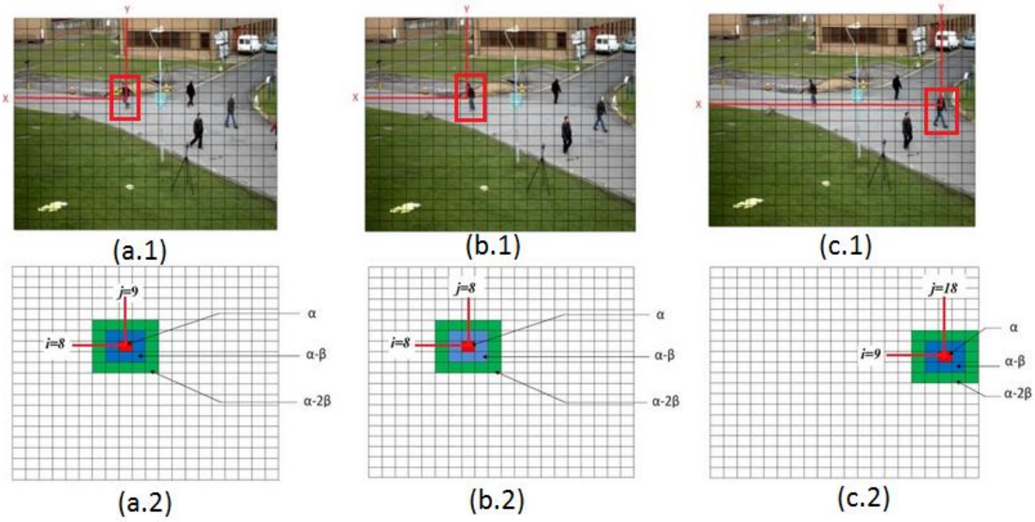

Fig. 7. Different Q position matrices. The red colored cell corresponding to the center of the object contains value α. (a.2), (b.2) and (c.2) show the position matrices of the detected objects.

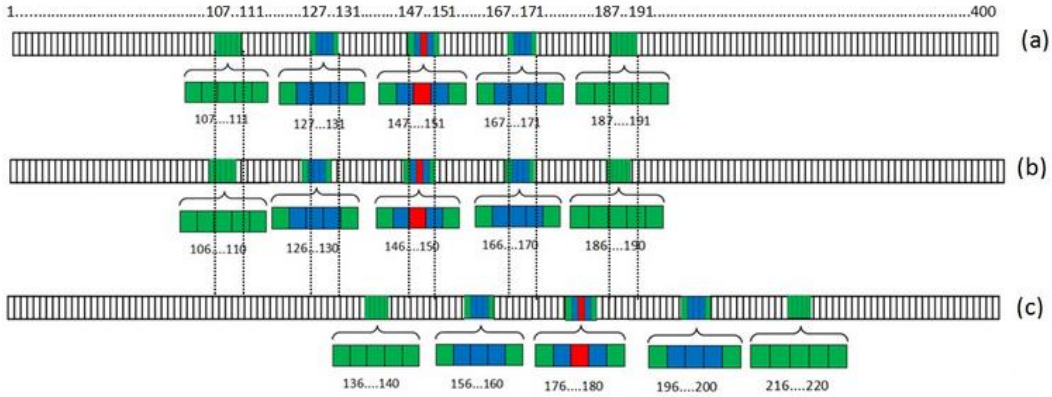

Fig. 8. Position vectors. (a), (b) and (c) represent the linearization of the position matrices corresponding to figures 7. (a.2), 7. (b.2) and 7. (c.2), respectively.

We call $Vposition \in \mathbb{R}^{m\times n}$ the vector obtained after a zig-zag scan of the Q position matrix. Let $E_t = \{1, \dots, P_t\}$ and $E_{t+1} = \{1, \dots, P_{t+1}\}$ be the set of indices of the objects detected at t and $t+1$, and $Vposition^l$ the position vector of object l . If $\langle Vposition^{\hat{\imath}}, Vposition^{\hat{u}}\rangle \gg 0$, Where $\hat{\imath} \in E_t$ and $\hat{u} \in E_{t+1}$, then the two objects are located in positions close to each other. If $\langle Vposition^{\hat{\imath}}, Vposition^{\hat{u}}\rangle \cong 0$, then the distance between the positions of both objects is large.

For example, we suppose that $Vposition^1$, $Vposition^2$, $Vposition^3$ are the position vectors shown in figures 8.a, 8.b and 8.c, respectively. Since the location of the colored areas of vector $Vposition^1$ matches that of vector $Vposition^2$ and does not match that of vector $Vposition^3$, the inner product $\langle Vposition^1, Vposition^2 \rangle \gg \langle Vposition^1, Vposition^3 \rangle$.

D. Overlapping Objects

As mentioned above, overlapping represents a tough problem in the detection process, especially when the background subtraction is used as the detection technique. The overlapped objects may be considered as one; this case is detected when the width of the bounding box containing the object exceeds 55% of its height. In the case where overlapping is lateral, we use the mask to divide the box into two boxes, each one containing the shape of one object. The distribution of white pixels in the upper part of the mask, permits to determine the number of overlapped objects (Fig.9). Threshold Δ is set in order to get the number of heads of people figured in the bounding box. Once the objects are correctly detected, the proposed tracker computes their appearance features again.

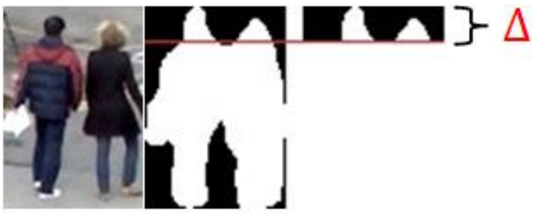

Fig. 9. Overlapped objects.

E. Object Descriptor

In this study, the proposed descriptor is defined as a concatenation of two major vectors: the appearance feature and the position feature. As introduced previously, the RGB histogram is built after dividing the body into two parts, the upper part and lower one from the center. Three vectors are obtained for each half, meaning six vectors to describe the entire body: $\{VR^{up} \in \mathbb{R}^{k\times1}, VG^{up} \in \mathbb{R}^{k\times1}, VB^{up} \in \mathbb{R}^{k\times1}, VR_{low} \in \mathbb{R}^{k\times1}, VG_{low} \in \mathbb{R}^{k\times1}, VB_{low} \in \mathbb{R}^{k\times1}\}$ where k represents the number of levels of each color channel. We denote by VH the concatenation of these six vectors. VH represents the vector encoding the appearance feature. The descriptor y of a detected object (see Fig.10), is a concatenation of seven vectors $VR^{up}, VG^{up}, VB^{up}, VR_{low}, VG_{low}, VB_{low}, and\ Vposition.$ Descriptor y is normalized in order to be used in the classification system employed for object recognition presented in section (IV).

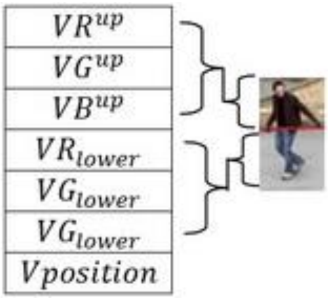

Fig.10. Object descriptor.

IV. Object Recognition Using Sparse Representation

Recognition can be formalized as a classification process through a linear regression model. This problem can be addressed by solving the ℓ^1 minimization based on sparse representation [8]:

$$y = A\beta + \varepsilon \tag{5}$$

Solving the linear system (5) can be the key to addressing the recognition problem; where $y \in \mathbb{R}^{M\times1}$ is the descriptor of the moving object and ε is the tolerated error. This vector contains information representing the object description defined above. All descriptors of objects detected in the previous frames are collected in a dictionary $A \in \mathbb{R}^{M\times N}$. The coefficients of $\beta \in \mathbb{R}^{\mathrm{N}}$ are needed to determine the class of this object.

A. Sparse Representation-based Classification (SRC)

The main goal of this work is to discriminate detected objects and to index them with the same label wherever they are located in the scene, even when they leave the scene and return back after a few frames. The used classification system based on sparse representation, can be expressed as follows:

The dictionary A, contains N object descriptors collected from previous frames.We denote by no the number of labeled objects, for the i^{th} one ,we regroup its n_i descriptors detected in the previous frames to create the object class $A_{ci} = [v_{i1}, v_{i2}, v_{i3}, \dots, v_{in_i}] \in \mathbb{R}^{M\times n_i}$, where $v \in \mathbb{R}^{M\times1}$, $0 < i \le no$ and $\sum_{i=1}^{no} n_i = N$. By proceeding this way, the dictionary can be considered as the concatenation of no classes corresponding to no labeled objects:

$A = [A_{c1}, A_{c2}, A_{c3}, \dots, A_{c\,no}]$. The descriptor y of the test object detected from current frame, can be approximated by the linear combination of known objects as in formula (5), where β =(β_1, β_2,..... β_{N}) $\in \mathbb{R}^{\mathrm{N}}$ is the sparse vector. It is stated *k-sparse* if $|supp(\beta)| \le k$, where $supp(\beta) = \{j: \beta_j \neq 0\}$

Fig.11.a, shows an input signal y to be linearly approximated. Figures 11.b, 11.c, 11.d and 11.e, represent classes of labeled objects. These classes constitute dictionary A. They contain descriptors of objects detected at the previous frames. As can be seen in Fig.11.b, the class contains three vectors representing the same object at different detection steps.

The sparse solution can be solved as a linear regression problem, and can be formulated by the following stable ℓ^1 -minimization problem derived from the sparse solution of formula (5) :

$$(\ell^1): \hat{\beta} = argmin\{\|y - A\beta\|_2 + \tau\|\beta\|_1\} \tag{6}$$

where τ is a cost parameter. To classify each new test object represented by y, we compute its sparse representation by minimizing the number of nonzero entries of sparse vector β and the residual.

$$\hat{\beta} = argmin\|\beta\|_1 \text{ subject to } \|y - A\beta\|_2 \le \varepsilon \quad (7)$$

Now, for associating the object represented by y to the i^{th} class in the training set, y can be approximated by:

$$\hat{y}_i = A\beta_i \quad (8)$$

where $\beta_i = [0,..,0,\beta_{i1},\beta_{i2},..,\beta_{in_i},0,..,0]$ is a vector obtained by setting to zero all entries not associated to the i^{th} class. The linear representation of y can be rewritten as:

$$\hat{y}_i = v_{i1}\beta_{i1} + v_{i2}\beta_{i2} + \cdots + v_{in_i}\beta_{in_i} \quad (9)$$

Using this approximation, any detected object can be assigned to the i^{th} class that satisfies:

$$class(y) = argmin_i\|y - \hat{y}_i)\|_2 = argmin_i\|y - A\beta_i)\|_2 \quad (10)$$

where $0 < i \le no$.

When non-zero coefficients are scattered over all classes, and:

$$\max\left(\sum_{l=1}^{n_i}\beta_{il}\right) < \theta\sum_{j=1}^{N}\beta_j \quad (11)$$

where $0 < \theta < 1$ is a threshold constant, then the object is considered as a new class.

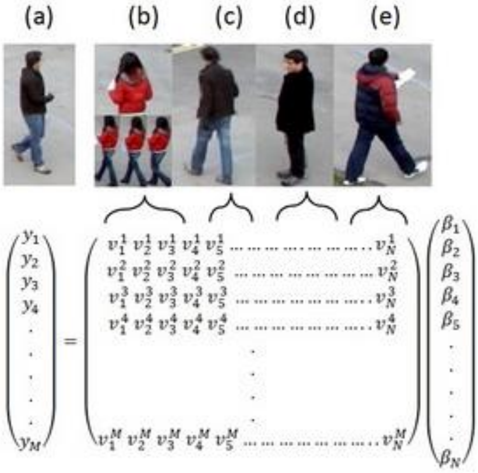
(a)
(b)
(c)
(d)
(e)
y_1 y_2 y_3 y_4 . . . . y_M
=
v_1^1 v_2^1 v_3^1 v_4^1 v_5^1 ... v_N^1
v_1^2 v_2^2 v_3^2 v_4^2 v_5^2 ... v_N^2
v_1^3 v_2^3 v_3^3 v_4^3 v_5^3 ... v_N^3
v_1^4 v_2^4 v_3^4 v_4^4 v_5^4 ... v_N^4
v_1^M v_2^M v_3^M v_4^M v_5^M ... v_N^M
β_1 β_2 β_3 β_4 β_5 . . . . β_N

Fig. 11. Graphic representation of the linear system y = Aβ. (a): Detected object test. (b), (c) , (d) and (e) represent classes of labeled objects.

Fig.12 represents an example of the adopted solution. The dictionary employed for the recognition contains 30 labeled training objects corresponding to 6 classes (Fig.12.a). After computing the sparse solution for two detected objects, the coefficients of the corresponding β vectors are represented in figures 12.b and 12.c. The first object has already appeared and thus exists in the dictionary. Its sparse coefficients are grouped in the second class (see the second class of Fig.12.a). However, the object in Fig.12.c is new, so it does not exist in the dictionary: as a consequence, its sparse coefficients are scattered over many classes.

As explained before, the problem of recognition consists of finding the values of the sparse coefficients. To get these values, SRC can operate as an iterative greedy algorithm that selects at each step the column of A which generates the largest inner product with the current residuals respecting formula (9). This is done by the OMP algorithm described below.

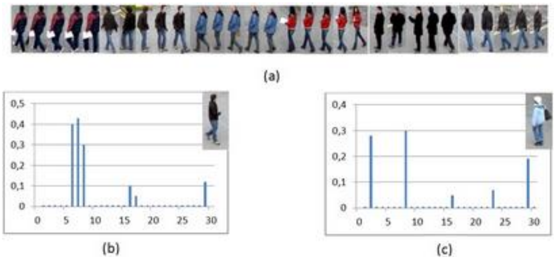

Fig. 12. Linear approximation of two objects, existing one and new one. (a) 30 labeled training objects corresponding to 6 classes. (b) Graph representing the sparse coefficients of a detected existing object. (c) Graph representing scattered coefficients of a new object.

B. OMP Algorithm

In SRC, several algorithms are proposed [30,41]. Our implementation consists of an OMP used to find the sparse coefficients. This algorithm was chosen for its speed of convergence due to the orthogonal projection [42,43]. In [15], Weizhi Lu et al. have shown that even when the dictionary has some false training samples, the sparse coefficients can be correctly grouped using the OMP algorithm.

The OMP algorithm can be presented as follows:

As inputs, we have a signal $y \in \mathbb{R}^{M\times 1}$ and a matrix $A \in \mathbb{R}^{M\times N}$ containing N columns. $A(c)$is a submatrix of A where $c \subseteq \{1,2,\dots,N\}$ represents a subset. We denote by t the number of iterations and by c_t the subset corresponding to the t^{th} iteration. Once the residual r_t or the number of iteration reaches some fixed threshold, then the vector $\hat{\beta} \in \mathbb{R}^{t\times 1}$ contains the sparse solution, where t is the last step.

```
Algorithm. 1: Orthogonal Matching Pursuit algorithm
Input: $y$ and $A$.
Output: $\hat{\beta}$.
Initialize: $r_0$=y, A($c_0$)={}.
  for t = 1; t = t + 1 do
      Stage 1. Find the column $A_i$ that solves the
      maximization problem    $argmax_i\langle A_i, r_{t-1}\rangle$
      where $i \in \{1,2,\dots,N\} \setminus c_{t-1}$.
      Update $c_t = c_{t-1} \cup \{i\}$.
      Stage 2. Compute $\hat{\beta} = \left(A(c_t)^T A(c_t)\right)^{-1} A(c_t)^T$
      and $P_t = A(c_t)\hat{\beta}$, where $P_t$ denote the projection
      onto the linear space spanned by the elements of
      $A(c_t)$.
      Stage 3. Update $r_t = (I - P_t)y$
      Stage 4. If $\|r_t\|_2 < \epsilon$ or $t$ reaches the maximum
      number of the authorized iterations, then stop the
  algorithm.
  end for.
```

C. Online Dictionary Updating

Once the object is classified, it will be stored temporarily until the end of the procedure for all objects. After that, it is added to the dictionary, more precisely in the appropriate class. All descriptors belonging to the same class will also be amended by updating the position part to become the same as the position part of the new descriptor added to their class.

V. Experiments And Results

The proposed algorithm is experimentally validated on a benchmark video and compared with other methods of the literature: MTUSR [15], ETHZ [44] and EPFL [45]. In our experiments the Q matrix is of size 20 x 20; it is set relatively to the size of the image and the moving objects. We computed the histogram and stored it without any quantification for more precision. So the size of each of the six histogram vectors is 256. The total size of our descriptor is therefore (256*6+20*20)=1936.

A. Dataset

We benchmarked our algorithm on PETS '09 dataset [46]. This video is filmed to be a reference in object tracking and it is used in many approaches. This video is not very crowded but the discrimination is a big challenge because the people wear clothes with similar colors. In 800 frames forming the video, 10 people must be initially indexed at their first appearance in the scene and must be recovered when they leave the scene and reappear again.

B. Results

All persons are correctly initialized and recognized [47]. They are indexed starting from 0 according to the order of their appearance in the video. The use of the position vector presented in section III-b, contributes when the colors are similar. In Fig.13.a the descriptor used for this example is built only with color histograms. When some objects are hidden (person number 2 on the left frame of Fig.13.a is hidden on the right frame), there is some confusion between this person and another who appears for the first time on the right frame of Fig.13.a. Fig.13.b is an example showing the efficiency of adding the position vector in our descriptor. Furthermore, the position vector makes the tracker robust in the case of the partial occlusion (see Fig.14). In Fig.15 the proposed method dealt successfully with overlapping and allowed detecting as much as possible moving objects. In frame_0160 of Fig.15, two objects are not detected by MTUSR because of overlapping, but they are well detected by ETHZ, EPFL and the proposed method. In frame_0031 of the same figure, two persons are considered as one by MTUSR, but are well separated by ETHZ, EPFL and our method. The reason is that, in the MTUSR approach, when the bounding box size containing the moving object is excessively large, the algorithm simply avoids this case of detection or considers two objects as one if they overlap a little; while we solved this kind of problem as explained in section III-d. The similarity of the clothes colors causes a confusion between intersecting objects. Some methods are not very robust to handle these situations. As can be seen in Fig.16, the overlapping reverses the indices of objects 6 and 8 between frame_0667 and frame_0675 using the EPFL method based on the trajectories of the moving objects; it is mentioned in [45] that EPFL is not applicable when the object's movements are way too erratic. The ETHZ, MTUHZ methods and the proposed approach based on the online trained dictionary, seem robust in these cases.

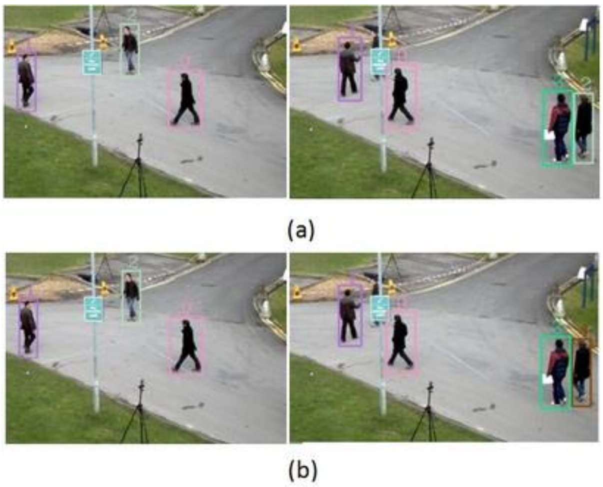

Fig. 13. Recognition using our algorithm.(a) New person is indexed as an existing person. (b) Correct initialization of new objects.

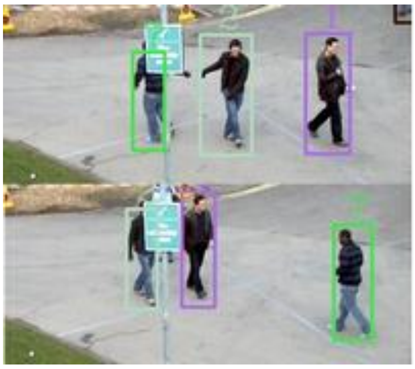

Fig. 14. Example of the partial occlusion.

When the objects leave the scene and reappear again in the future frames, the system must recover them. In Fig.17 our method, maintains the index of the person number 5 in frame_0063 and frame_0471. The MTUSR approach works well too. Both MTUHZ and our method use the appearance features and the position feature; but for ETHZ using a particle filtering framework and for the EPFL methods, we can see that the recovery process is false, and the same person is differently colored in the frames. In Fig.18, the persons indexed as 3 and 4 in frame_0031 are well recovered in frame_0535 and frame_0626 by the proposed method. MTUSR does not recover the persons indexed as 4 and 5 in frame_0535 and they are considered as a new person indexed as 10. In frame_0626, only the person indexed as 4 is recovered by MTUSR. ETHZ does not detect the person in frame_0535 and switches the identity of two persons in frame_0626. EPFL does not recover the persons correctly between frames_0031 and the other frames. The reason making our method more effective is the way to represent the objects as described in section III-b.

Table 2 summarizes the number of the persons entries and reappearance in the scene. The proposed approach seems robust for a good recognition. According to Table 2, MTUSR, ETHZ and EPFL have respectively 2, 6 and 7 errors of initialization and recovery.

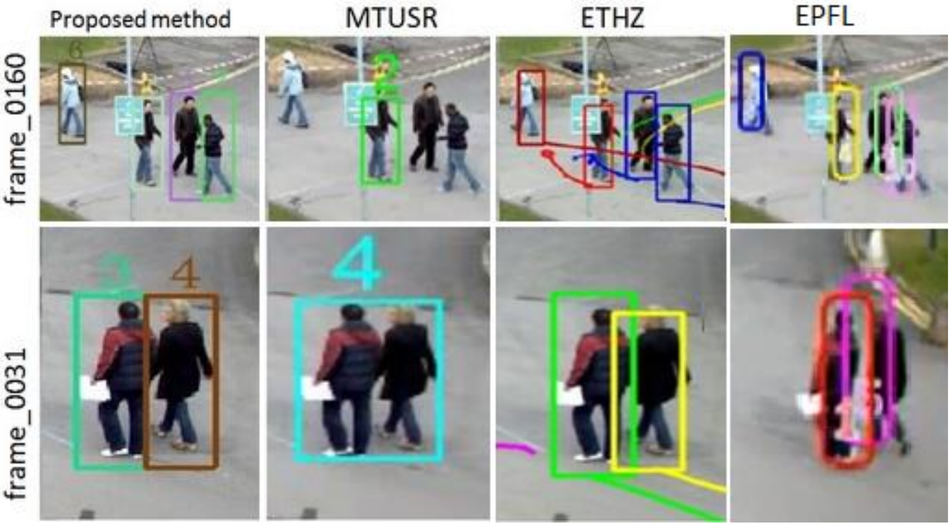

Fig. 15. Visual comparison of the overlapping treatment with four approaches.

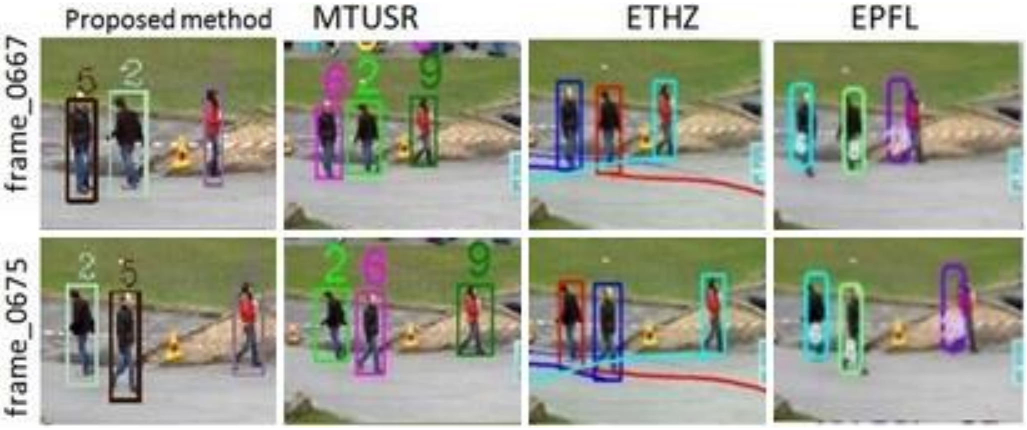

Fig. 16. Illustration of the problem of identity switch caused by overlapping: visual comparison of our method with MTUSR, ETHZ and EPFL.

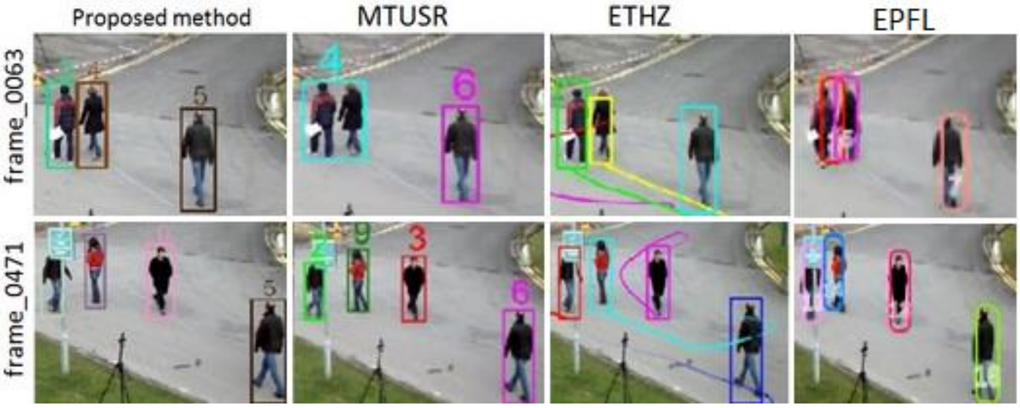

Fig. 17. Illustration of the recovery process with a visual comparison of the proposed method with the MTUSR, ETHZ and EPFL recovery processes.

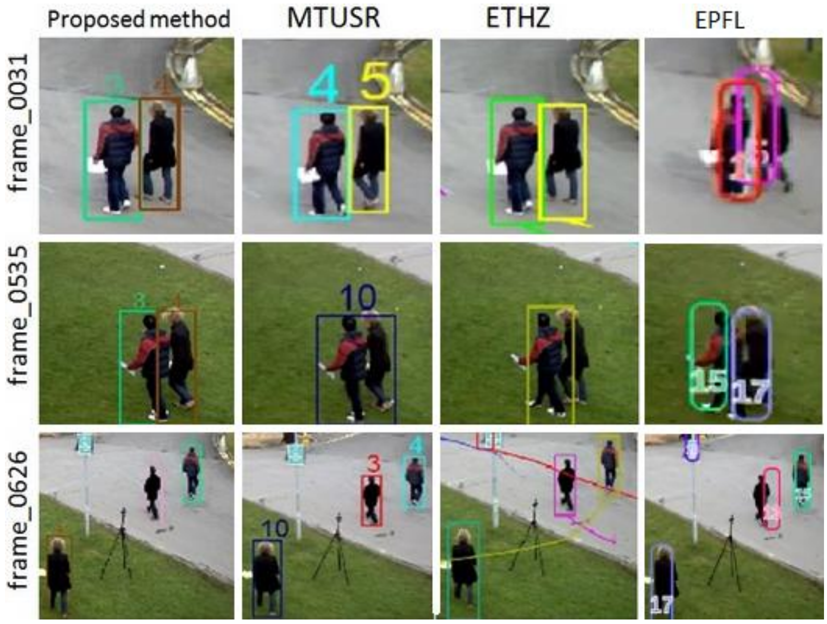

Fig. 18. Illustration of the problem of the false initialization and the recovery process with a visual comparison of the proposed method with MTUSR, ETHZ and EPFL.

Table 2. Objects entries and reappearance in the scene, the F value means false object initialization. The last column contains the number of errors of each method

Objects	0	1	2	3	4	5	6	7	8	9	**N**
Number of entries	2	4	2	2	2	2	1	1	2	1	**E**
Proposed approach	2	4	2	2	2	2	1	1	2	1	**0**
MTUSR	2	4	2	1	1	2	1	1	2	1	**2**
ETHZ	2	4	2	1	1	1	1	1	F	F	**6**
EPFL	1	2	1	1	1	1	1	1	2	1	**7**

VI. Conclusion

In this paper, a tracking algorithm based on sparse representation with an online dictionary learning is proposed. The proposed tracker uses only the color histogram and position as information to describe the moving object. To assure the persistent identity of the objects, we have proposed a new way to describe the object position and to improve its appearance feature. Experimental results obtained by applying this new technique to a complex benchmark video, demonstrate the efficiency of our approach compared to some existing methods. To get this precision, we have built large descriptors and thus the processing time has increased with the growth of the dictionary. We therefore suggest as a perspective work the use of some parallel implementation to reduce the computation time.

Acknowledgments

This work was supported in part by the Franco-Algerian cooperation program PHC TASSILI grants. Entitled “Système Conjoint de Compression et d’Indexation Basé-Objet pour la Vidéo (SCCIBOV),” the project gathers members of Institut National des Sciences Appliqu ées (INSA) de Rennes, Universit é de Technologie de Compi ègne (UTC), the University of Sidi Bel Abb ès and the University of Mascara.

References

[1] J. Starck, M. Elad, and D. Donoho, “Image decomposition via the combination of sparse representation and a variational approach,” IEEE Trans. on Image Processing, vol. 14, no. 10,pp. 1570–1582, 2005.

[2] Y. Li, A. Cichocki, and S. Amari, “Analysis of sparse representation and blind source separation,” Neural Computation, vol. 16, no. 6, pp. 1193–1234, 2004.

[3] B. Olshausen, P. Sallee, and M. Lewicki, “Learning sparse image codes using a wavelet pyramid architecture,” in NIPS, 2001, pp. 887–893.

[4] M. Elad and M. Aharon, “Image denoising via learned dictionaries and sparse representation,” in CVPR, 2006.

[5] M. Elad, B. Matalon, and M. Zibulevsky, “Image denoising with shrinkage and redundant representation,” in CVPR, 2006.

[6] S. Zhao,Z.Hu " Modular Fisher Discriminant Sparse Representation for robust face recognition " Optik 125 (2014) 6505–6508

[7] E. Cande`s, “Compressive Sampling,” Proc. Int’l Congress of Mathematicians, 2006.

[8] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” IEEE Tran. PAMI, vol. 31, pp. 210–227, 2009.

[9] M. Yang and L. Zhang, “Gabor feature based sparse representation for face recognition with gabor occlusion dictionary,” in ECCV, 2010.

[10] S. Chen, D. Donoho, and M. Saunders, “Atomic decomposition by basis pursuit,” SIAM J. Scientific Computing , vol. 20, no. 1, pp. 33–61, 19.

[11] I. Drori and D. Donoho, “Solution of L1 minimization problems by LARS/Homotopy methods,” in ICASSP, 2006, vol. 3, pp. 636–639.

[12] Emmanuel J. Candès and Yaniv Plan. "Near-ideal model selection by ℓ1 minimization". Ann. Statist. Volume 37, Number 5A (2009), 2145-2177.

[13] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,” in 27th Annual Asilomar Conference on Signals, Systems, and Computers, 1993.

[14] S. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” IEEE Trans. on Signal Processing, vol. 41, pp. 3397–3415, 1993.

[15] W. Lu, C. Bai, K. Kpalma and J. Ronsin "Multi-object Tracking using Sparse Representation", ICASSP 2013, May 26-31, 2013, Vancouver, Canada (2013).

[16] M. Andriluka, S. Roth, and B. Schiele, “People-Tracking-by-Detection and People Detection-by-Tracking,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008.

[17] J. Berclaz, F. Fleuret, and P. Fua, “Robust People Tracking with Global Trajectory Optimization,” Proc. IEEE CS Conf. Computer Vision and Pattern Recognition, 2006.

[18] H. Fang, J. Kim and J. Jang "A Fast Snake Algorithm for Tracking Multiple Objects" Journal of Information Processing Systems, Vol.7, No.3, September 201

[19] B. Leibe, K. Schindler, N. Cornelis, and L.V. Gool, “Coupled Object Detection and Tracking from Static Cameras and Moving Vehicles,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 30, no. 10, pp. 1683-1698, Oct. 2008.

[20] M. J. Black. Eigentracking: Robust matching and tracking of articulated objects using a view based representation. In ECCV, pages 329–342, 1996.

[21] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi. Robust online appearance models for visual tracking. PAMI, 25(10):1296–1311, 2003.

[22] D. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. IJCV, 77(1-3):125–141, 2008.

[23] S. Avidan. Support vector tracking. In CVPR, pages 184–191, 2001.

[24] S. Avidan. Ensemble tracking. PAMI, 29(2):261–271, 2007.

[25] J. Shao et al. " Multi-part sparse representation in random crowded scenes tracking "Pattern Recognition Letters 34 (2013) 780–788

[26] M. Xue and L. Haibin, “Robust visual tracking using l1 minimization,”in Proc. Int. Conf. Comput. Vision, 2009, pp. 1436–1443.

[27] C. Bao, Y. Wu, H. Ling, and H. Ji, “Real time robust l1 tracker using accelerated proximal gradient approach,” in Proc. IEEE Conf. Comput.Vision Pattern Recogn., Jun. 2012, pp. 1830–1837.

[28] Z. Han, J. Jiao, B. Zhang, Q. Ye, and J. Liu, “Visual object tracking via sampled-based adaptive sparse representation,” Pattern Recogn., vol. 44, no. 9, pp. 2170–2183, Mar. 2011.

[29] Q. Wang, F. Chen, W. Xu, M. Yang : "Online discriminative object tracking with local sparse representation". WACV 2012: 425-432

[30] B. Efron, T. Hastie, and R. Tibshirani, “Least angle regression,” Annals of Statistics, vol. 32, pp. 407–499,2004.

[31] L. Wang, H. Yan, K. Lv, and C. Pan "Visual Tracking Via Kernel Sparse Representation With Multikernel Fusion", IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 24, NO. 7, JULY 2014.

[32] Y. Lin, Q. Yu, G. Medioni "Efficient detection and tracking of moving objectsin geo-coordinates" Machine Vision and Applications (2011) 22:505–520.

[33] W. Sai, L. Zhengyi, H. Chuanwen "An Implementation of Distance Measure for Dynamic Scene" 2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP) June 9 – 11, 2013, Beijing, China.

[34] A. Amato, M. Mozerov, F. X. Roca, and J. Gonzlez,“Robust real-time background subtraction based on local neighborhood patterns,” EURASIP J. Adv. Sig. Proc., 2010.

[35] Yu, Guoshen . "Solving Inverse Problems with Piecewise Linear Estimators: From Gaussian Mixture Models to Structured Sparsity". *IEEE Transactions on Image Processing* 21 (5): 2481–2499.

[36] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[37] P. Viola and M. Jones, "Robust real-time face detection," presented at Second International Workshop on Statistical and Computational Theories of Vision - Modeling, Learning, Computing, and Sampling, Vancouver, Canada, pp. 1-25, 2001.

[38] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," presented at 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. I-511-I-518 vol.1, 2001

[39] A. Belgherbi.,A. Bessaid: " Morphological Segmentation of the Spleen From Abdominal CT Images " IJIGSP Vol.4, No.4, May 2012

[40] Liang JI, Jim Piper, "Erosion and dilation of binary images by arbitrary structuring elements using interval coding", Pattern Recognition Letters Volume 9, Issue 3, April 1989, Pages 201–209

[41] Q. Feng , "Novel classification rule of two-phase test sample sparse representation" Optik 125 (2014) 5825–5832

[42] T. Tony Cai and Lie Wang "Orthogonal Matching Pursuit for Sparse Signal Recovery With Noise" IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 57, NO. 7, JULY 2011

[43] J. Zhang , H. Zhang , Z. Li ,"A hierarchical structure with improved OMP sparse representation used with face recognition", Optik 125 (2014) 4729–4735

[44] M. D. Breitenstein, F. Reichlin, B. Leibe, E. KollerMeier, and L. V. Gool, "Online multi-person tracking-by-detection from a sing and uncalibrated camera," IEEE Trans. PAMI, vol. 33, no. 9, pp. 1820–1833, 2011.

[45] H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua, "Tracking multiple people under global appearance constraints,"in ICCV, 2011.

[46] J. Ferryman, A.-L. Ellis "Performance evaluation of crowd image analysis using the PETS2009 dataset" Pattern Recognition Letters 44 (2014) 3–15

[47] "Online Multi-object tracking using Sparse representation " https://www.youtube.com/watch?v=8WX2Rvn36mQ.

Authors' Profiles

Mohamed ELBAHRI received the Dipl.Ing. in computer science and Magister degree in electronic departement from Djillali Liabes University, Sidi Bel-Abbes, Algeria, and is currently a Ph.D. candidate in computer science. His principle research interests are in the fields of computer vision and image processing.

Kidiyo KPALMA received his PhD in Image Processing from the National Institute of Applied Sciences of Rennes (INSA) in 1992 and his HDR (Habilitation à diriger des recherches) in Signal processing and Telecommunications from the University of Rennes 1 in 2009 in France. He is currently Full Professor at INSA where he teaches signals and systems, signal processing and DSP. As a member of the Image and Automatic department of the Institute of Electronics and Telecommunications of Rennes (IETR), his research interests are image analysis, pattern recognition, image segmentation, image fusion and remote sensing.

Nasreddine TALEB received the M.Sc. degree in computer engineering from Boston University, Boston, MA, USA, the Electrical Engineering degree from Northeastern University, Boston, MA, USA, and the Ph.D. degree in electrical engineering from Djillali Liabes University, Sidi Bel-Abbes, Algeria. He is currently a Professor with the Department of Electronic Engineering, University of Djillali Liabes, where he has been teaching since 1990 and where he is also a Senior Research Scientist and the Director of the "Communication Networks, Architecture, and Multimedia" Laboratory. His principle research interests are in the fields of digital signal and image processing, image analysis, medical and satellite image applications, pattern recognition, and advanced architectures for the implementation of DSP/DIP applications

Miloud CHIKR El-MEZOUAR received the Dipl.Ing. and Magister degrees in electrical engineering from Djillali Liabes University, Sidi Bel-Abbes, Algeria, and the Ph.D. degree in signal and image processing, under joint supervision, from Djillali Liabes University and from the Institut National des Sciences Appliquées (INSA), Rennes, France. He joined as a Lecturer in Djillali Liabes University in 2002 and is currently an Associate Professor with the Department of Electronic Engineering at the same University. His principle research interests are in the fields of digital signal and image processing, image analysis, medical and satellite image applications, remote sensing and pattern recognition.

How to cite this paper: Mohamed ELBAHRI, Kidiyo KPALMA, Nasreddine TALEB, Miloud CHIKR EL-MEZOUAR,"A Novel Object Position Coding for Multi-Object Tracking using Sparse Representation", IJIGSP, vol.7, no.8, pp.1-12, 2015.DOI: 10.5815/ijigsp.2015.08.01