

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
31 January 2008 (31.01.2008)

PCT

(10) International Publication Number
WO 2008/012631 A1

(51) International Patent Classification:
G06T 7/20 (2006.01)

(21) International Application Number:
PCT/IB2007/002050

(22) International Filing Date: 20 July 2007 (20.07.2007)

(25) Filing Language: Italian

(26) Publication Language: English

(30) Priority Data:
TO2006A000556 27 July 2006 (27.07.2006) IT

(71) Applicant (for all designated States except US):
VIDEOTEC S.P.A. [IT/IT]; Via Friuli, 6, I-36015 Schio (VI) (IT).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **GENNARI, Giambattista** [IT/IT]; c/o Videotec S.p.A., Via Friuli 6, I-36015 Schio (VI) (IT). **RACCANELLI, Giorgio** [IT/IT]; c/o Videotec S.p.A., Via Friuli 6, I-36015 Schio (VI) (IT). **FREZZA, Ruggero** [IT/IT]; c/o Università degli Studi di Padova, Dipartimento di Ingegneria dell'Informazione, Via Gradenigo 6/B, I-35131 Padova (IT). **CENEDESE, Angelo** [IT/IT]; c/o Università degli Studi di Padova, Dipartimento di Ingegneria dell'Informazione, Via Gradenigo

6/B, I-35131 Padova (IT). **CAMPANA, Enrico** [IT/IT]; Via Belluno 5, I-36016 Thiene (VI) (IT).

(74) Agent: **DINI, Roberto**; c/o Metroconsult S.r.l., Via Sestriere 100, I-10060 None (TO) (IT).

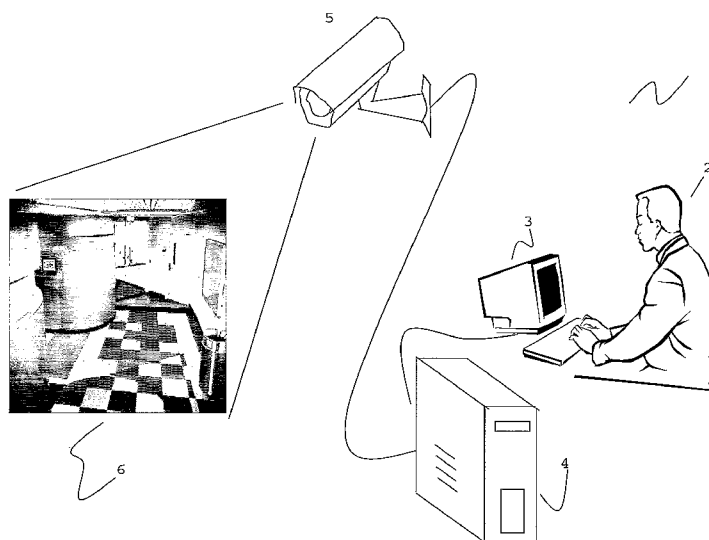
(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— with international search report

[Continued on next page]

(54) Title: EVENT DETECTION METHOD AND VIDEO SURVEILLANCE SYSTEM USING SAID METHOD



(57) Abstract: An event detection method for video surveillance systems and a related video surveillance system are described. The method comprises a learning phase, wherein learning images of a supervised area are acquired at different time instants in the absence of any detectable events, and an operating detection phase wherein current images of said area are acquired. The method detects an event by comparing a current image with an image corresponding to a linear combination of a plurality of reference images approximating, or coinciding with, respective learning images.

WO 2008/012631 A1



-
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*
- For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

EVENT DETECTION METHOD AND VIDEO SURVEILLANCE SYSTEM USING SAID METHOD

DESCRIPTION

Field of invention

The present invention relates to an event detection method according to the preamble of claim 1 and to a video surveillance system using said method.

Description of the prior art

In the present patent, the term “video surveillance system” refers to a surveillance system using at least one image acquisition unit and being capable of acquiring sequences of images of a supervised area.

Generally, known video surveillance systems include event detection systems which can generate alarms when an anomalous event takes place.

Some known systems only detect variations (beyond a certain user-defined threshold) in the brightness of the pixels of two successive images, such as two frames of a video signal or two images taken by a camera at different times.

These systems suffer from the drawback that sudden light variations due, for example, to travelling clouds or reflections on water, or natural movements within the supervised area (e.g. the branches of a tree moved by the wind or authorized cars driven down a street) may cause several false alarms.

For overcoming these drawbacks, video surveillance systems are known which comprise a learning phase wherein the system builds a model of the supervised area in a normal situation, i.e. a situation wherein no alarm should be triggered. During the operating phase, the pixels of the taken image are compared with the pixels of the model. If the difference in the pixels is beyond a certain operator-defined threshold, an alarm will be triggered.

Notwithstanding the creation of a model pertaining to a normal situation of the supervised area, the effectiveness of a pixel-by-pixel comparison between the acquired image and the model is often poor because a single pixel differing from its model is sufficient to trigger an alarm.

This leads to the generation of a lot of false alarms.

The problem of the generation of false alarms has been addressed by some known solutions (American patent US 5,892,856), which de facto subtract from the detection those pixels that show intensity variations (e.g. due to natural movements in the watched scene) during the

learning phase.

Such a solution, which is not very effective, can only be used in selected contexts such as presence detection at workstations (as in patent US 5,892,856).

Other more advanced solutions, such as the one disclosed by patent application US 2004/0246336, provide in primis for the creation of a statistic model for each pixel (with related average and variance); subsequently, during the event detection phase, the system extrapolates the image of the detected object/person in order to compare it with a set of models of authorized objects.

However, this solution has the drawback that it requires much available memory for storing the models of the scene and of the authorized objects, as well as high computing power for analyzing the whole image in real time by comparing the detected objects with the authorized ones.

A problem which is common to all known solutions is due to the fact that the alarm is triggered when the brightness or colour difference of a pixel in two consecutive frames is beyond a certain operator-defined threshold. This results in the efficiency of the system being dependent on the operator's skill, this being a problem if the operator is not an expert.

The main object of the present invention is to overcome the drawbacks of the prior art, and in particular to provide a video surveillance system and an event detection method which allow for a more effective detection of events while reducing the number of false alarms and preferably while not requiring high capacity in terms of available memory and computing power.

It is a further object of the present invention to provide a system having a high degree of automation and being capable of calculating automatically the error threshold beyond which a normal variation of one pixel must be discriminated from an alarm condition.

The present invention also aims at providing a video surveillance system and an event detection method capable of optimizing memory usage and of varying the computing complexity depending on the dynamics being present in the supervised area during the learning phase.

These and further objects of the present invention are achieved through a video surveillance system and method according to the appended claims, which are intended as an integral part of the present description.

Brief description of the invention

The present invention is based on the idea of leaving the traditional approach according to

which every single pixel of a current image is compared with a respective pixel of a reference image or with a pixel model.

In particular, the invention aims at taking into consideration regions of the image, i.e. groups of pixels, in order to take also into account the correlation among the pixels when detecting an event, thus reducing the number of false alarms.

The video surveillance system according to the invention acquires images of a supervised area and compares single regions thereof with respective "models" representing a normal situation, which are built in the form of a space of images acquired during a learning phase, said images relating to a normal situation of the watched scene.

The image or region is treated like an image vector, the difference of which from a normal situation is measured as a projection error of the image vector on a space of images representing the "model" of the supervised area in a normal situation.

The "model" is built by starting from a set of images acquired during a learning phase by shooting the area in a normal situation.

The learning phase may include a model validation phase substantially consisting in a simulation of an operating detection phase. The validation phase uses images of the scene in a normal situation acquired during the learning phase, and checks whether the model just built is good or not.

For detecting events, the method according to the invention exploits in particular the properties of principal components analysis (PCA).

Advantageously, the method also provides for a suitable reduction of the informative content of the acquired images, thus ignoring minor phenomena occurring in a scene and reducing the number of false alarms.

Brief description of the drawings

Further objects and advantages of the present invention will become apparent from the following description and from the annexed drawings, wherein:

- Fig. 1 shows a video surveillance system according to an embodiment of the invention;
- Fig. 2 is a block diagram of the processing applied to the acquired images by a video surveillance system according to the invention;
- Fig. 3 shows an acquired image broken down into a plurality of regions.
- Fig. 4 shows an example of event detection.

Detailed description of the invention

Fig. 1 shows a video surveillance system 1. Through a monitor 3, an operator 2 watches the

images 6 acquired by an image acquisition unit 5.

In Fig. 1, the image acquisition unit is a video camera capable of providing an output video, i.e. a continuous sequence of images, but it is understood that, for the purposes of the invention, it may be replaced with any other equivalent means, e.g. a programmed digital camera acquiring images at regular time intervals.

An image may thus correspond to a frame or a half-frame of a video signal acquired by the video camera, to a static image acquired by a digital camera, to the output of a CCD sensor, or more in general to a portion of the above.

As well known, a digital or analog image can be disassembled into pixels, i.e. fundamental elements of the image.

One or several matrixes may therefore be associated with each image, the elements of which are the voltage values of the analog video signal or the brightness or colour values of the pixels.

When a colour video camera is used, the acquired image will correspond to a tridimensional matrix wherein each matrix element (i.e. each pixel) corresponds to a triplet of values corresponding to the values of the RGB signals.

In a greyscale image, each matrix element is associated with a value corresponding to the grey value of the corresponding pixel.

In the present description this image-matrix association will be implicit, so that reference will be made below, for example, to rows and columns of an image.

Back to Fig. 1, the video camera 5 shoots an area, in this specific case a corridor, and transmits a video signal which can be displayed on the monitor 3.

According to the invention, the surveillance system 1 includes an image processing unit 4 capable of detecting events starting from the images acquired by the video camera 5.

In Fig. 1, the image processing unit 4 is represented by an electronic computer connected to the video camera 5 and to the monitor 3 in order to receive and process the video signal sent by the video camera and to display images on the monitor.

In a preferred embodiment, the image processing unit 4 is a video server, i.e. a numerical computer receiving a video signal from the image acquisition unit, processing it according to the method of the present invention, and transferring a video signal to one or several terminals connected thereto, said terminal being in particular an operator's workstation.

Of course, many other solutions are possible as well, e.g. the image processing unit 4 may be incorporated in the video camera 5 (which in such a case will comprise an image acquisition

unit and an image processing unit), which will be connected to the monitor 3 either directly or through a video switch.

The image processing unit 4 is provided with software containing code portions capable of implementing the event detection method described below.

According to said method, a learning phase is executed at least once at installation time, wherein the system builds a “model” of the watched scene in a normal situation.

Nevertheless, according to said method the learning phase may advantageously be repeated several times under different environmental conditions (light, traffic, etc.). This allows to build one or several models.

For example, several models may be built at different times of the day, in which case an image acquired at a certain time shall be compared with the valid model for that time.

In a normal situation, i.e. in the absence of any detectable events, the operator 2 starts the software learning phase, wherein images of the supervised area are acquired which will be hereafter referred to as “learning images”.

In the preferred embodiment described below, the acquired images correspond to the frames of the video signal generated by the video camera 5.

During this phase, it is possible that moving objects are taken, such as leaves of trees or vehicles travelling down a street behind the scene, so that the acquired images may differ from one another.

Consequently, the model can represent the watched scene in a dynamic situation, without any events to be detected.

As shown in Fig. 2a, the first step of the learning phase of the event detection method consists in the selection (202) of a set of N frames $F_1, \dots, F_N$ starting from the video signal 201 acquired by the video camera. Preferably, these frames are subjected to image processing operations such as a greyscale conversion (203) in order to reduce the size of the data to be treated, and possibly, additionally or alternatively, a low-pass filtering (204) with a Gaussian kernel in order to eliminate and smooth any high-frequency variations not to be detected, thus reducing the informative content of the images to be treated and focusing the detection on the interesting informative content of the image.

The frames thus modified are then inserted into a learning buffer (205).

As an alternative, the above image processing steps may be repeated cyclically on each acquired frame, as shown in Fig. 2b. In this case, the parameter n is set initially to 1 and an image is acquired (202b), which is then converted to greyscale (203b), filtered with a low-pass

filter (204c) and stored in the buffer (205c). Subsequently, the value of n is incremented and these steps are repeated until N images are stored in the buffer.

Once created, the content of the learning buffer is subdivided into two parts: a first group of frames, called “training frames”, on which a principal components analysis (PCA) is carried out, and a second group of frames, called “validation frames”, used for validating the results obtained from the PCA.

Therefore, the learning phase correspondingly comprises a training phase and a validation phase.

According to a preferred embodiment, during the training phase each one of the training frames F_n (the number of which is S in this example), and preferably each one of the frames stored in the learning buffer, is subdivided by means of a predefined grid into regions (i.e. small images preferably square or rectangular in shape) $R_{i,j}$ having max. $M \times M = m$ pixels.

The result, as shown in Fig. 3, is a plurality of portions of images obtained from each frame.

The size of the grid depends on the typical dimensions of the target to be discovered in the watched scene.

Said grid may then be set up by an installer at installation time depending on the shooting perspective and on the operator's needs, or else be predefined at the factory.

For each region $R_{i,j}$ of a frame F_n , a corresponding column vector $IR_{i,j}(F_n)$ is obtained.

This vector $IR_{i,j}(F_n)$ is substantially obtained by progressively entering the elements of the matrix $R_{i,j}$ (i.e. the values of the pixels of the region), which meet together when scrolling the columns from the top and from the left. Thus, the element $IR_{i,j}(F_n)(2)$ corresponds to the pixel located on the second row of the first column of the image $R_{i,j}$.

By placing the column vectors of a same region $R_{i,j}$ side by side, a corresponding normality matrix $Y_{i,j} = (IR_{i,j}(F_1), IR_{i,j}(F_2), \dots, IR_{i,j}(F_S))$ is created.

The columns of the normality matrix generate a vectorial space of the images.

The columns carry the information relating to one region of the watched scene at different instants and in a normal situation, whereas the autovectors of the respective co-variance matrix are the principal components thereof, i.e. the directions in which the variance of the columns of $Y_{i,j}$, i.e. of the collected images, is greater.

Once the matrix $Y_{i,j}$ has been obtained, a singular value decomposition (SVD) is carried out in order to obtain three matrixes $U_{i,j}$, $V_{i,j}$, $\Sigma_{i,j}$ such that

$$Y_{i,j} = U_{i,j} \cdot \Sigma_{i,j} \cdot V_{i,j}^T$$

with

$$\begin{aligned} U_{ij} &= [u_1 \dots u_s] \\ \Sigma_{ij} &= \text{diag}(\sigma_1, \dots, \sigma_s) \\ V_{ij} &= [v_1 \dots v_s] \end{aligned}$$

wherein u_i (vectors $\in \mathbb{R}^m$) are the autovectors of the co-variance matrix of Y_{ij} , and $\sigma_1, \dots, \sigma_s$ are the singular values of Y_{ij} .

As known, when using an SVD decomposition the elements of the diagonal matrix Σ_{ij} are bound by the following relationship:

$$\sigma_1 \geq \dots \geq \sigma_r \geq \sigma_{r+1} \geq \dots \geq \sigma_s \geq 0$$

According to the invention, in order to optimize the detection of events and to focus on the relevant informative content of the image, the matrix Y_{ij} is approximated by the matrix

$$Y_{ij}^r = U_{ij}^r \cdot \Sigma_{ij}^r \cdot (V_{ij}^r)^T$$

wherein

$$\begin{aligned} U_{ij}^r &= [u_1 \dots u_r] \\ \Sigma_{ij}^r &= \text{diag}(\sigma_1 \dots \sigma_r) \\ V_{ij}^r &= [v_1 \dots v_r] \end{aligned}$$

and wherein the number of singular values r taken into consideration for building Y_{ij}^r is obtained by ignoring all singular values below a certain threshold.

The matrix Y_{ij}^r has the same dimensions as the matrix Y_{ij} , but it only carries the information relating to the first r principal components of the matrix Y_{ij} .

The columns $u_1 \dots u_r$ of the matrix U_{ij}^r are the principal components of the matrix Y_{ij} .

To determine the threshold, several tests have been carried out which have shown that a good event detection can be achieved when the informative content of Y_{ij} is approximated by giving up 20%-30%, preferably 25%, of the energy of Y_{ij} (i.e. of the image portions R_{ij} used for building this matrix).

Due to a known property related to singular values, the percentage of energy $\%E(r)$ bound to the first r principal components of the matrix Y_{ij} (with S singular values) is:

$$\%E(r) = \frac{\sum_{k=1}^r \sigma_k^2}{\sum_{k=1}^S \sigma_k^2}$$

Starting from these considerations, for each region R_{ij} a respective value of r is determined and the matrix Y_{ij}^r is built, which represents the essence of what was learnt during the

learning phase.

At this point the validation phase is carried out, aiming at verifying that the learning set consisting of the training frames is sufficiently representative of the watched scene in a normal situation.

The verification provides a simulation of an operating detection phase, wherein current images of the supervised area are replaced with at least one validation frame F^{VAL} , i.e. a learning image not belonging to the learning set, and therefore not used for building the normality matrix Y_{ij} .

In practice, the validation is carried out by using at least one validation frame F^{VAL} which is subdivided into a plurality of regions $R_{ij}(F^{VAL})$ through the same grid already used for subdividing the training frames.

For each region $R_{ij}(F^{VAL})$, a corresponding vector $IR_{ij}(F^{VAL})$ is created as was previously done for the training frames.

Subsequently, the vector $IR_{ij}(F^{VAL})$ is projected on a space of the learning images in order to determine the “distance” between the validation image and the normal situation synthesized in the matrix Y_{ij}^r .

As known, given a matrix $Y_{i,j}^r = (y_1^r, y_2^r \dots y_s^r)$ with elements in $\mathfrak{R}^m$, it is possible to define the space $\text{Range}(Y_{i,j}^r)$, which is an subspace of $\mathfrak{R}^m$, consisting of all linear combinations of the columns $y_1^r, y_2^r \dots y_s^r$. According to a known linear algebra theorem, the subspace $\text{Range}(Y_{i,j}^r)$ coincides with the subspace $\text{Range}(U_{i,j}^r)$, consisting of all linear combinations of the columns of $U_{i,j}^r$, i.e. of the first r principal components of $Y_{i,j}$.

The projection of an image portion R_{ij} of the frame F^{VAL} , i.e. of the vector $IR_{ij} \in \mathfrak{R}^m$, on the principal components of Y_{ij} is therefore obtained by using the projector operator defined as

$$P_{\text{Range}(Y_{i,j}^r)} = U_{i,j}^r \cdot (U_{i,j}^r)^T$$

Once this operator has been calculated, for each image portion R_{ij} the projection $\text{Proj}(IR_{i,j})$ and the projection error $\text{err\_Proj}(IR_{i,j})$ are calculated as

$$\begin{aligned} \text{Proj}(IR_{i,j}) &= U_{i,j}^r \cdot (U_{i,j}^r)^T \cdot IR_{i,j} \\ \text{err\_Proj}(IR_{i,j}) &= \|\text{Proj}(IR_{i,j}) - IR_{i,j}\|_2 \end{aligned}$$

The watched scene will be signalled as anomalous (i.e. an event will be detected) if the

projection error is greater than the respective threshold, i.e. if the following relationship is fulfilled:

$$\text{err\_Proj}(\text{IR}_{i,j}) \geq \text{Thr}_{i,j}$$

According to a preferred embodiment, the threshold $\text{Thr}_{i,j}$ is determined automatically and is set to the $r+1^{\text{th}}$ singular value σ_{r+1} of the matrix $Y_{i,j}$, i.e. to the highest index singular value fulfilling the relationships

$$\sum_{i=1}^r \sigma_i \leq \%E \quad \text{and} \quad \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq \sigma_{r+1} \geq 0$$

Advantageously, the threshold can be set to $k\sigma_{r+1}$ with $k>1$ so as to take into account any background noise being present in the images.

Since the validation frames F^{VAL} should represent the scene in a normal situation, no event should be detected even in the worst conditions. Otherwise, the set of acquired training frames is not representative of the scene in a normal situation; according to the method, it will be necessary in this case to select a different learning set and to rebuild the matrix $Y_{i,j}^r$.

For the purpose of preventing a good set of training frames from being discarded because a validation frame (not the training frames) is not actually representative of a normal situation, in a preferred and advantageous embodiment the method provides for simulating the event detection on a plurality of validation frames.

According to this embodiment of the method, the set of training frames will be changed if a number of events is detected which is greater than a preset percentage, e.g. 25%, calculated on the total number of measurements. For example, if the validation phase verifies 100 validation frames and detects 25 events, then the method will require the learning set to be rebuilt.

According to a further advantageous embodiment, the normality matrix $Y_{i,j}$ will be regenerated by starting from a new learning set if the mean projection error $\overline{\text{err\_Proj}}$ of a plurality of validation images is greater than or equal to the threshold $\text{Thr}_{i,j}$.

The mean projection error $\overline{\text{err\_Proj}}$ is calculated as

$$\overline{\text{err\_Proj}} = \text{media}[\text{err\_Proj1}, \text{err\_Proj2}, \text{err\_Proj3}, \dots]$$

wherein “media” is a function which receives the projection errors err_Proj1 , err_Proj2 , err_Proj3 ... of the validation images and outputs the mean value thereof.

In another embodiment it is possible to calculate the maximum projection error $\overline{\text{err\_Proj}}_{\text{MAX}}$ among the projection errors of a plurality of validation images. If said error is beyond a preset threshold, then the training phase will be repeated by rebuilding a new learning set.

In a preferred embodiment, the number of frames to be added depends on the difference between the mean projection error $\overline{\text{err\_Proj}}$ and the threshold.

As an alternative, a predetermined number of frames may be added which is dependent on the memory available in the buffer.

Preferably, not all of the frames added to the learning set are consecutive.

The maximum size of the learning set may be defined (at installation or production time) as a function of the available memory and computing power.

During the learning phase, the method provides for storing a plurality of pieces of information which are useful for detecting an event. In particular, the following information is stored:

- the matrixes $U_{i,j}^r = [u_1 \dots u_r]$
- the thresholds $\text{Thr}_{i,j}$
- the number r of principal components of each region $R_{i,j}$ which allows to optimize the matrix $Y_{i,j}$.

Once the learning phase is over, the video surveillance system can start the actual operating phase by acquiring current images of the supervised area in order to detect any events.

Through the video camera 5, a current image of the supervised area is acquired, in particular a frame F^* of the video signal generated by the video camera.

The frame F^* is subdivided into a plurality of regions $R_{i,j}(F^*)$ by means of the same grid previously used during the learning phase.

For each region $R_{i,j}(F^*)$, a corresponding vector $IR_{i,j}(F^*)$ is created, which is then projected on the principal components of $Y_{i,j}$, i.e. on the space of the learning images, in order to evaluate the projection error

$$\text{err\_Proj}(IR_{i,j}) = \|\text{Proj}(IR_{i,j}) - IR_{i,j}\|_2.$$

as described with reference to the validation phase.

Within each region $R_{i,j}$, an event will be detected if the projection error is greater than the threshold selected by the system for that same region $R_{i,j}$.

These steps use the information stored during the learning phase.

Fig. 4 illustrates an example of event detection. The analysed frame F^* shows a corridor, which is the same corridor as that shown in Figs. 1 and 3, wherein two people are present whose silhouettes occupy a portion of the frame; in the example of Fig. 4, said silhouettes occupy seven regions $R_{i,j}$.

Only these regions $R_{i,j}$ (highlighted by a white contour) occupied by people will have an error which is greater than the accepted threshold, thus triggering an alarm.

Although the above-described algorithm provides for treating images as vectors, obviously the physical meaning of these operations should not be overlooked, such operations being also obtainable in different manners and through different steps.

Detecting an event by evaluating the projection error of the image vector on the vectorial space of the images $\text{Range}(Y_{i,j}^T)$ corresponds to comparing a current image $R_{i,j}(F^*)$ with an image corresponding to a linear combination of a plurality of reference images approximating, or coinciding with, respective learning images.

This concept will become clearer when looking at Fig. 5.

The space $\text{Range}(Y_{i,j}^T)$ is represented by the plane Γ in which the reference images $y_1^T \dots y_S^T$ and all linear combinations $C_1, C_2, C_3 \dots$ of such reference images lay.

Projecting a current image $R_{i,j}(F^*)$ on the plane Γ means comparing said image with the linear combination C_i of the reference images which is closer to the image $R_{i,j}(F^*)$ in the sense of the L2 standard.

Rather than a pixel-by-pixel comparison, this is an image-by-image comparison; this means that, in addition to the values of the pixels of the image, the correlation thereof is also taken into account through the principal components analysis.

Of course this comparison, which according to the preferred and advantageous embodiment of the present invention is synthesized by the step of projecting the vector corresponding to the current image on the space of the images $\text{Range}(Y_{i,j}^T)$, may alternatively be obtained by comparing first the values of the single pixels and then the relationships among the pixels of a current image and of an image built during the learning phase and corresponding to a linear combination of images approximating, or coinciding with, images acquired in a normal situation.

The advantages of the present invention are apparent from the above description.

The proposed approach, based on a comparison between portions of the current image (i.e. groups of pixels) and respective models built during the learning phase, turns out to be more reliable than known solutions based on a comparison of each pixel, in that it allows to take also into account the spatial correlation among the pixels.

Furthermore, unlike known solutions using a statistic model in order to represent the normal situation of one pixel, the model according to the invention consists of a set of vectorial spaces, each pertaining to one region of the image, obtained from a learning phase.

As a result, therefore, the model according to the invention allows to represent normality in a much more accurate manner than the solutions known in the art, since it uses images rather

than statistical parameters such as average and variance.

Also, PCA is an information compression method which allows to rebuild the original data (this not being possible with the statistic methods used by existing solutions), thus not endangering the quality with which normality is described and the resulting detection reliability.

The proposed approach also offers a high degree of automation, unlike traditional approaches which require arbitrary choices when programming and setting up the system.

In particular, the detection threshold is calculated automatically based on the learning data.

Moreover, the proposed approach selects the length of the learning buffer autonomously depending on the dynamics being present in the watched scene during the learning phase. In particular, the algorithm chooses a compromise solution between available memory and scene dynamics.

Furthermore, the minimum number of principal components used for representing the vectorial space also changes as a function of the scene dynamics and is also calculated by the algorithm.

It is also clear that the above description presents a video surveillance system and an event detection method according to preferred embodiments which represent non-limiting examples of realization of the system and of implementation of the method.

In particular, the selection of the optimum number of components r_{ij} , which allows to reduce the informative content of the matrix Y_{ij} , can be carried out in several different manners.

For example, the r_{ij} principal components may be chosen by picking up those associated with a singular value being greater than a percentage (preferably in the range of 1-4%, more preferably 2%) of the greatest singular value (σ_1) of the matrix Y_{ij} .

For example, if the matrix Y_{ij} has the following singular values: $\sigma_1=100$, $\sigma_2=20$, $\sigma_3=10$, $\sigma_4=1$, $\sigma_5=0,5$, $\sigma_6=0,3$..., by accepting singular values greater than 2% of σ_1 only the first three principal components u_1 , u_2 , u_3 will be taken into consideration.

Furthermore, the projection error, which in the preferred embodiment is calculated as $\text{err\_Proj}(\text{IR}_{ij}) = \|\text{Proj}(\text{IR}_{ij}) - \text{IR}_{ij}\|_2$, may without distinction be calculated as any L_p standard:

$$\text{err\_Proj}(\text{IR}_{ij}) = \|\text{Proj}(\text{IR}_{ij}) - \text{IR}_{ij}\|_{L_p}$$

CLAIMS

1. Event detection method for video surveillance systems, comprising a learning phase, wherein learning images of a supervised area are acquired at different times in the absence of any detectable events, and an operating detection phase, wherein current images of said area are acquired, characterized in that an event is detected by comparing a current image with an image corresponding to a linear combination of a plurality of reference images approximating, or coinciding with, respective learning images.
2. Method according to claim 1, characterized in that said reference images are obtained by using a method for analysing the principal components of a learning set consisting of a plurality of learning images.
3. Method according to claim 2, wherein said learning set is organised in a normality matrix, the columns of which contain the pixels of said learning images, and said reference images are the columns of a matrix (Y_{ij}^r) approximating said normality matrix (Y_{ij}) .
4. Method according to claim 3, characterized in that said method of principal components analysis comprises a step of decomposing said normality matrix into singular values, this decomposition providing three matrixes

$$\begin{aligned} U_{ij} &= [u_1 \dots u_s], \\ \Sigma_{ij} &= \text{diag}(\sigma_1, \dots, \sigma_s), \text{ and} \\ V_{ij} &= [v_1 \dots v_s], \end{aligned}$$

such that the normality matrix Y_{ij} can be written as:

$$Y_{ij} = U_{ij} \cdot \Sigma_{ij} \cdot V_{ij}^T$$

5. Method according to claim 4, wherein said matrix approximating said normality matrix is

$$Y_{ij}^r = U_{ij}^r \cdot \Sigma_{ij}^r \cdot (V_{ij}^r)^T$$

wherein

$$\begin{aligned} U_{ij}^r &= [u_1 \dots u_r] \\ \Sigma_{ij}^r &= \text{diag}(\sigma_1 \dots \sigma_r) \\ V_{ij}^r &= [v_1 \dots v_r] \end{aligned}$$

the number r of columns of the matrixes $U_{ij}^r, \Sigma_{ij}^r, V_{ij}^r$ being smaller than the number S of the columns of the matrixes $U_{ij}, \Sigma_{ij}, V_{ij}$.

6. Method according to any of claims 3 to 5, wherein said matrix approximating the normality matrix is such that the sum of its singular values is smaller than a preset percentage, preferably 75%, of the sum of the singular values of the normality matrix.
7. Method according to any of claims 3 to 5, wherein the singular values of said matrix

approximating the normality matrix are greater than a preset percentage, preferably between 2% and 4%, of the greatest singular value of said normality matrix.

8. Method according to any of the preceding claims, characterized in that said comparison comprises the step of projecting an image vector, corresponding to said current image, on the vectorial space defined by all of the linear combinations of the reference images as a whole.

9. Method according to claim 8, characterized in that an event will be detected if the following relationship is fulfilled:

$$\text{err\_Proj(IR)} \geq \text{Thr}$$

wherein Thr is a threshold and err_Proj(IR) is a projection error calculated as the standard of the difference between the vectors IR and Proj(IR), IR being said image vector and Proj(IR) being the vector corresponding to the projection of IR on said vectorial space.

10. Method according to claim 9 when claim 8 is dependent on any of claims 2 to 7, wherein said normality matrix and said approximating matrix have a plurality of singular values in common, characterized in that said selected threshold is equal to the greatest non-common singular value.

11. Method according to claim 9 when claim 8 is dependent on any of claims 2 to 7, wherein said normality matrix and said approximating matrix have a plurality of singular values in common, characterized in that said selected threshold is equal to the greatest non-common singular value multiplied by a preset real number greater than one.

12. Method according to claim 10, characterized in that said selected threshold is equal to a preset percentage between 2% and 4% of the greatest singular value of said normality matrix.

13. Method according to any of the preceding claims, characterized in that said learning phase comprises a training phase and a validation phase, said training phase being adapted to select a learning set consisting of said respective learning images, and said validation phase being adapted to verify that said learning set is a model which is representative of said area in the absence of any detectable events.

14. Method according to claim 13, characterized in that said validation phase comprises at least one simulation step for simulating said operating phase, said simulation being carried out by comparing a validation image with a linear combination of said reference images, said validation image being a learning image not belonging to said learning set.

15. Method according to claim 13 or 14, characterized in that the training phase will be repeated by changing the learning set if said validation phase detects an event.

16. Method according to claim 13 or 14, wherein said validation phase comprises the steps of:

- projecting a plurality of image vectors, corresponding to a plurality of validation images, on the vectorial space defined by all of the linear combinations of said reference images as a whole,
- for each projection, detecting a projection error defined as the standard of the difference between an image vector and the vector corresponding to the projection of the image vector on said vectorial space.

17. Method according to claim 16, characterized by

- calculating the mean projection error $\overline{\text{err\_Proj}}$ as

$$\overline{\text{err\_Proj}} = \text{media}[\text{err\_Proj1}, \text{err\_Proj2}, \text{err\_Proj3}, \dots]$$

wherein “media” is a function receiving the projection errors err_Proj1 , err_Proj2 , err_Proj3 ... of the validation images and returning the mean value thereof,

- repeating the training phase by changing the learning set, if the mean projection error $\overline{\text{err\_Proj}}$ of a plurality of validation images is greater than or equal to a certain threshold.

18. Method according to claim 16, characterized by

- determining the maximum projection error $\overline{\text{err\_Proj}}_{\text{MAX}}$ of all projections of said plurality of image vectors,
- repeating the training phase by changing the learning set, if the maximum projection error $\overline{\text{err\_Proj}}_{\text{MAX}}$ is greater than or equal to a certain threshold.

19. Method according to claim 13 or 14, characterized in that the training phase will be repeated by changing the learning set, if said validation phase detects a percentage of events which is greater than a preset threshold.

20. Method according to any of the preceding claims, characterized in that image processing steps are carried out during both the learning phase and the event detection phase.

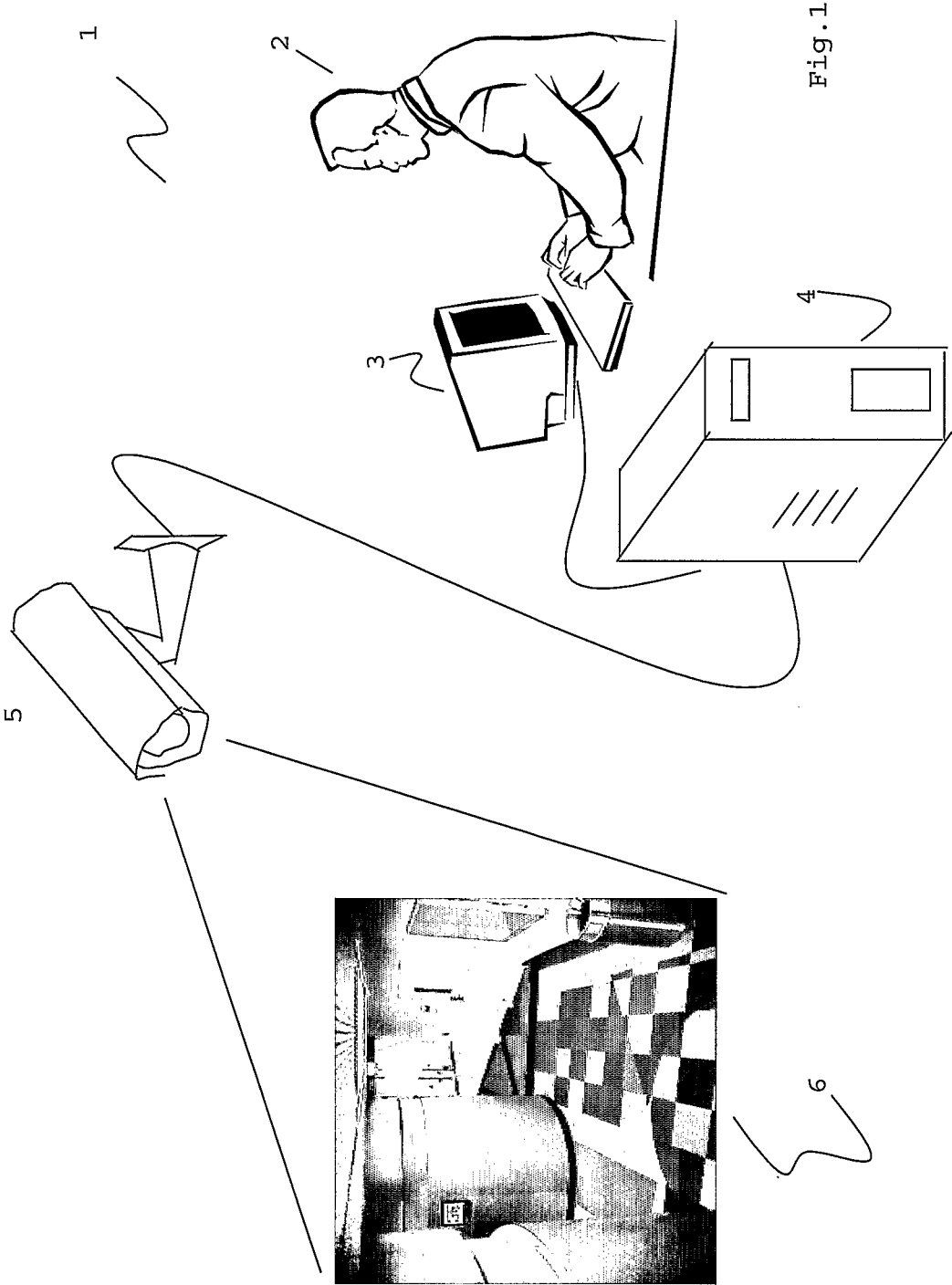
21. Method according to claim 20, wherein said image processing steps comprise a step for reducing said images to greyscale.

22. Method according to claim 20 or 21, wherein said image processing steps comprise a low-pass filtering step preferably using a Gaussian kernel.

23. Method according to any of the preceding claims, wherein said learning images and said current images are groups of pixels obtained by subdividing frames or half-frames of a video signal by means of a predefined grid.

24. Information technology product which can be loaded in a memory area of an electronic computer and which comprises code portions adapted to implement the method according to any of claims 1 to 23 when executed by said computer.

25. Video surveillance system comprising at least one image acquisition unit connected to an image processing unit, said image processing unit being adapted to implement at least a portion, in particular all, of the method according to any of claims 1 to 23.
26. Method according to the innovative teachings of the present description and of the annexed drawings, which show preferred and advantageous examples of embodiment of said method.
27. System according to the innovative teachings of the present description and of the annexed drawing, which show preferred and advantageous examples of realisation of said system.



2/4

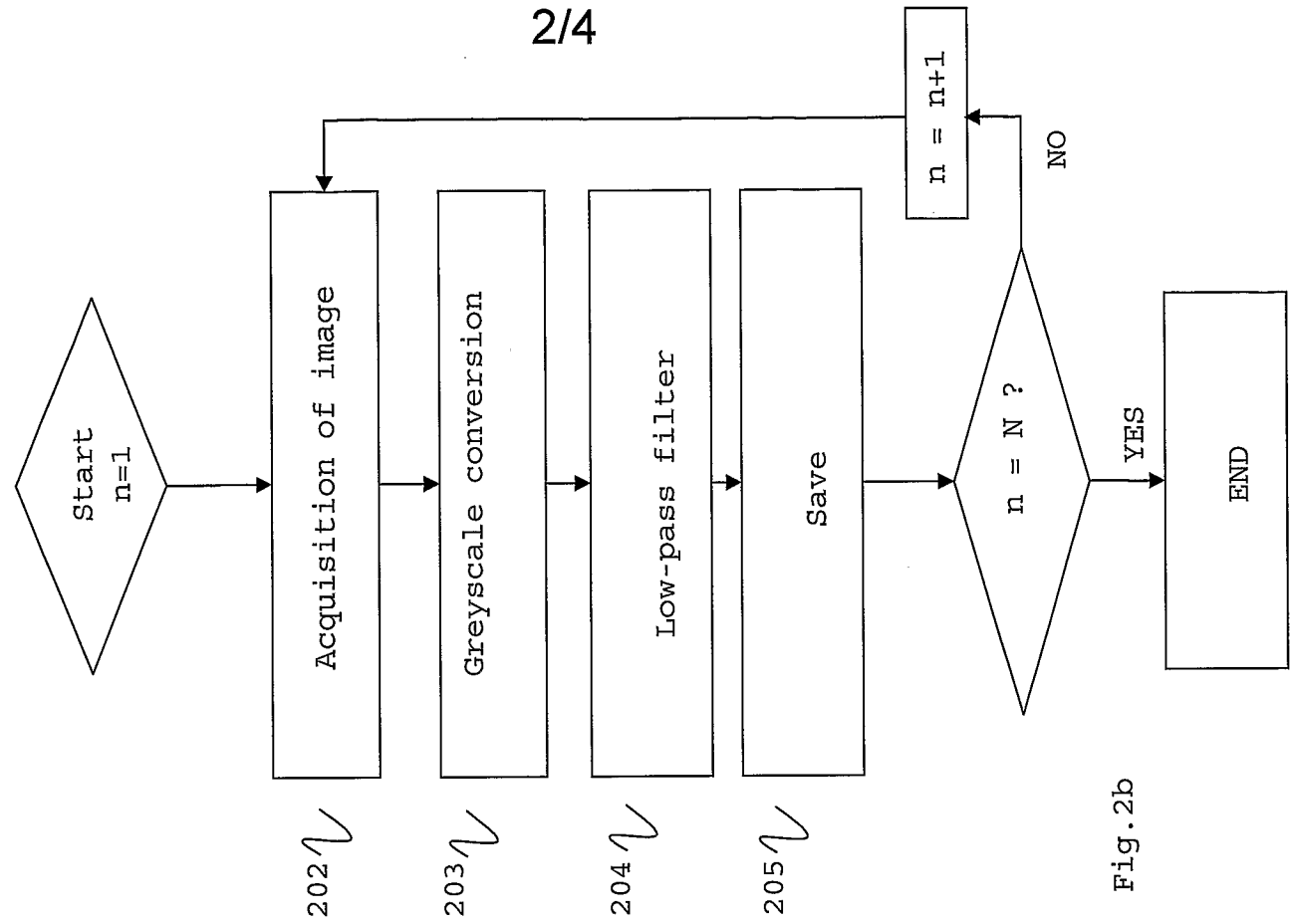


Fig. 2b

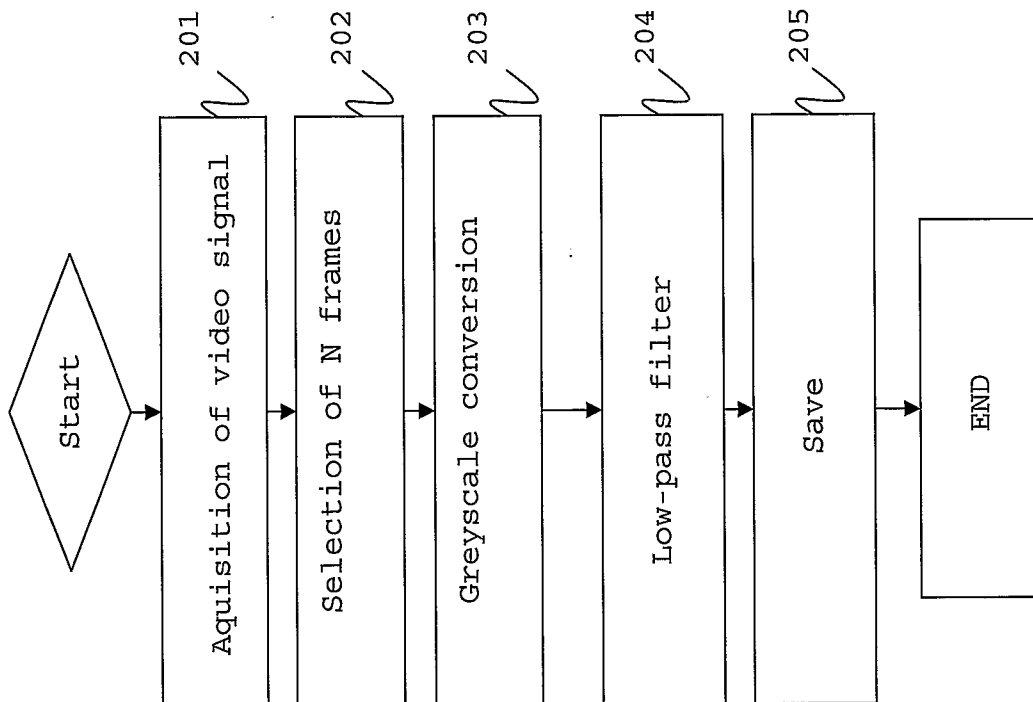


Fig. 2a

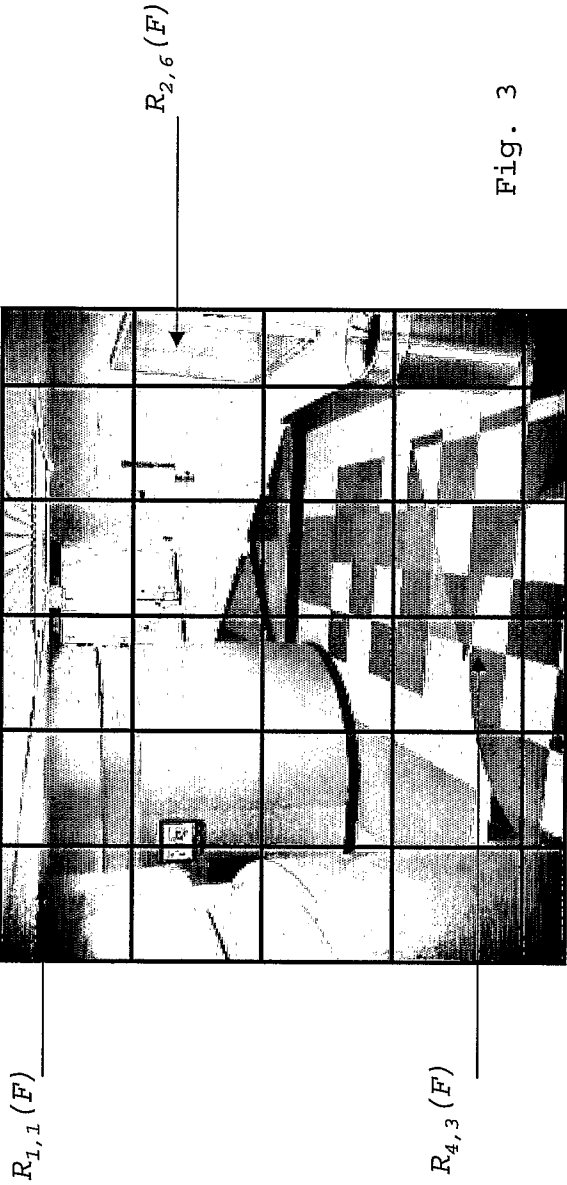


Fig. 3

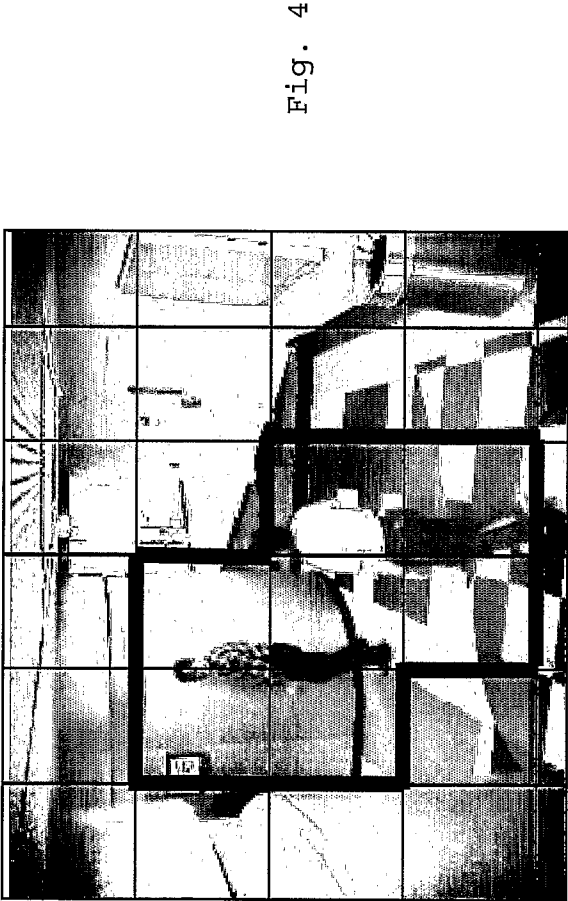


Fig. 4

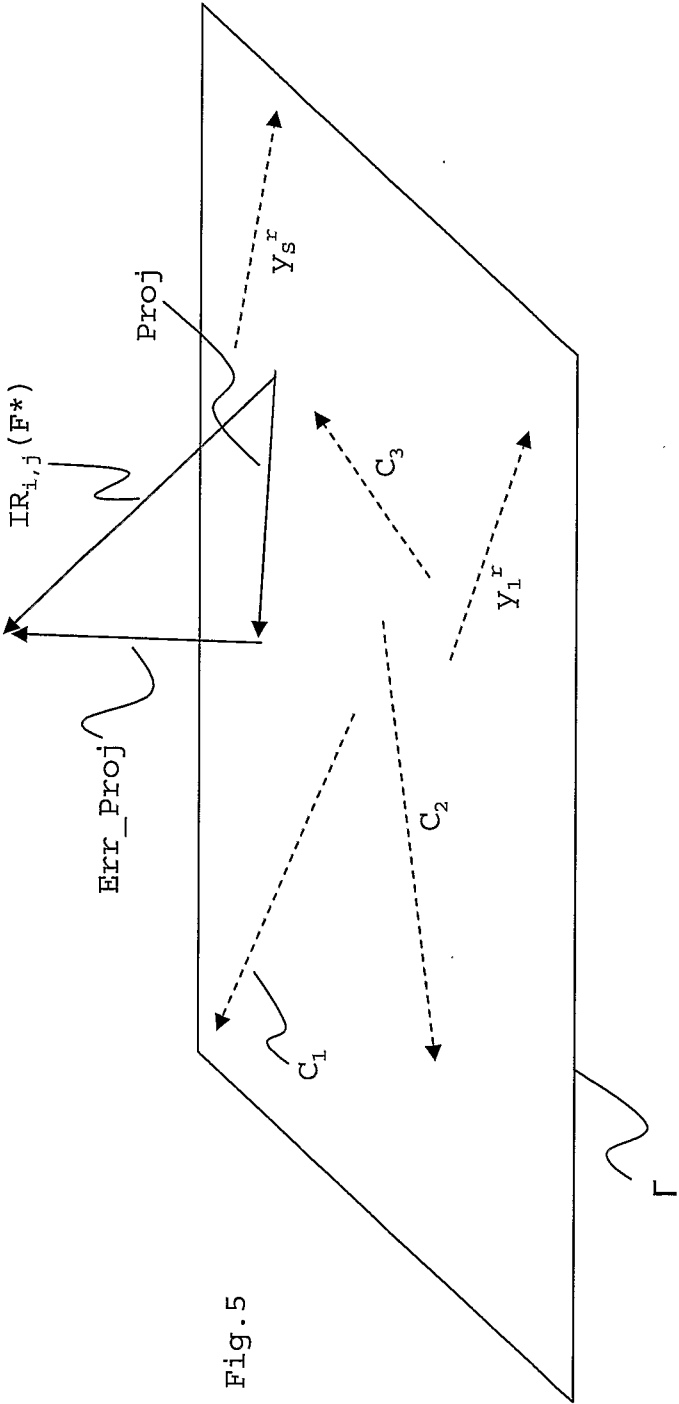


Fig. 5

INTERNATIONAL SEARCH REPORT

International application No

PCT/IB2007/002050

A. CLASSIFICATION OF SUBJECT MATTER

INV. G06T7/20

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06T G06K G08B

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, WPI Data, INSPEC, COMPENDEX

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	EP 1 542 154 A (SONY UK LTD [GB]) 15 June 2005 (2005-06-15) abstract paragraphs [0026] - [0034], [0108], [0109], [0208] pages 29-32 figure 17 ----- -/--	1-3, 8, 9, 20, 22-25 4-7, 10-16, 19, 21



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

- *A* document defining the general state of the art which is not considered to be of particular relevance
- *E* earlier document but published on or after the international filing date
- *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- *O* document referring to an oral disclosure, use, exhibition or other means
- *P* document published prior to the international filing date but later than the priority date claimed

- *T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- *X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- *Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
- *G* document member of the same patent family

Date of the actual completion of the international search

21 November 2007

Date of mailing of the international search report

06/12/2007

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Eckert, Lars

INTERNATIONAL SEARCH REPORT

International application No

PCT/IB2007/002050

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	YANWEI PANG ET AL: "Face Recognition Using Neighborhood Preserving Projections" ADVANCES IN MULTIMEDIA INFORMATION PROCESSING - PCM 2005 LECTURE NOTES IN COMPUTER SCIENCE;;LNCS, SPRINGER-VERLAG, BE, vol. 3768, 2005, pages 854-864, XP019024109 ISBN: 3-540-30040-6 sections "1 Introduction", "2 Overview of the Proposed Method: NPP", "3 Locally Linear Embedding (LLE)", "5 Face Recognition Using NPP", "6 Experimental Results" section "6.1 Experiment with the Yale Face Database"	1,2,8,9, 24
Y	HUAZHONG NING ET AL: "Silhouette analysis-based gait recognition for human identification" IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE, NEW YORK, NY, US, vol. 25, no. 12, December 2003 (2003-12), pages 1505-1518, XP011103919 ISSN: 0162-8828 abstract sections "1 Introduction", "1.1 Overview of Approach", "3 Features Extraction", "3.3.1 PCA Training", "3.3.2 Projection"	4-7, 10-12
Y	WEI FAN ET AL: "Locally Linear Models on Face Appearance Manifolds with Application to Dual-Subspace Based Classification" COMPUTER VISION AND PATTERN RECOGNITION, 2006 IEEE COMPUTER SOCIETY CONFERENCE ON NEW YORK, NY, USA 17-22 JUNE 2006, PISCATAWAY, NJ, USA, IEEE, 17 June 2006 (2006-06-17), pages 1384-1390, XP010923196 ISBN: 0-7695-2597-0 abstract sections "1 Introduction", "2 Previous Work", "4 Local Model Construction"	13-16, 19,21

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/IB2007/002050

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 1542154 A	15-06-2005	GB 2409029 A US 2005128306 A1	15-06-2005 16-06-2005