Performance Evaluation of Mobile Person Detection and Area Entry Tests through a One-View Camera

Abderrahmane Ezzahout[1], Youssef Hadi[1, 2], Rachid Oulad Haj Thami[1]
[1]RIITM Research Group ENSIAS
Mohamed-V University-Souissi
Rabat, Morocco
[2]Faculty of Sciences
Ibn Tofail University
Kenitra, Morocco
abderrahmane.ezzahout@um5s.net.ma, hadi-moulay-youssef@univ-ibntofail.ac.ma, oulad@ensias.ma

ABSTRACT: *In recent years, detection and tracking people from a video stream have been widely studied in many commercial and public surveillance systems. Moving object detection is considered as a crucial phase of automatic video surveillance systems. Particularly, people detection is the first important step in any technique of video tracking processes which can be divided into many stations such as motion estimation, tracking people, etc. Several methods have been developed for this problem of separating the foreground and background pixels in video surveillance. This paper focuses on computable evaluation of some people detection algorithms for four different video sequences. Our study is based on quantitative and qualitative results respectively by calculating the loss of foreground pixels. In this study, three methods have been evaluated by using two metrics: False Negative Error (FNE) and False Positive Error (FPE). In the result we choose the algorithm which minimizes the Error (%). Practically, the good technique which dominates on the video surveillance applications is the statistical representation of pixels in foreground which is named Gaussian Mixture Model (GMM). In the second part of this paper we control the people entering in a supervised region by detection of the first correct pixel incoming to our supervised area, and we trigger an alarm system.*

Keywords: Moving Object Detection, Background Subtraction, Surveillance System, Misclassified Pixels

Received: 29 June 2012, Revised 4 August 2012, Accepted 13 August 2012

1. Introduction

Over the past years, several surveillance techniques have matured dramatically. This progression provides organizations with significant opportunities to improve security and reduce operating costs. Today's, private and public agencies are faced with a serious need to safe clients and employees, improve productivity and increase profit as well as customer satisfaction with a security system. Operators are additionally challenged with running good quantities of information in various forms. Especially, captured files by camera in public environments.

Surveillance systems have been limited and passive for most of the industry's history. Building from years of experience in wide

field-of-view imaging and detection algorithms. The force of a surveillance system depends upon the consistency of its algorithms. The quality of a system's algorithms can make a difference between a threat being detected as a critical alarm. Enterprise, Organization, institution etc, have developed surveillance systems not only to help them to detect and respond to dangers sooner, but also to help them center of attention on improving business operations.

Video surveillance has typically involved the placement of cameras in strategic and sensitive areas, this serves not only as a deterrent to crime, but also to track the movement of people and property. The use of video surveillance systems results in the establishment of multiple of videos that must be viewed by security guards. The cost of employing security personnel to monitors hundreds of cameras, in addition to storing a high volume of video sequences can be too expensive and sequences can contain poor image quality and deteriorate over time.

Video surveillance systems typically use multiple video cameras, transmitting the video signals to a central control room, where a multiplex matrix is used to display some of the images to security personnel. Event detection and recognition requires the perceptual capabilities of human operators to detect and identify objects moving within the field-of-view (FOV) of the cameras and to appreciate their actions. No matter how vigilant the operators, manual monitoring inevitably suffers from information overload, as a result of periods of operator inattention due to fatigue, distractions and interruptions. In practice, it is inevitable that a significant proportion of the video channels are not regularly monitored, and potentially important events are overlooked. Moreover, tiredness increases dramatically as the number of cameras in the system is increased. Automating all or part of this process would provide significant benefits, ranging from the capability to alert an operator to these potential events of interest, during to a fully automatic detection and analysis system. However, the trustworthiness of automated detection systems is a very important issue, since frequent false alarms induce skepticism in the operators, who quickly learn to ignore the system.

As the name suggests, the process of background subtraction is to separate the foreground and the background pixels in a sequence of video frames. This paper analyzes three techniques for background subtraction, with minimal, medium and high degrees of complexity and gives the results of each of them by calculating the error estimation of the misclassified pixels from the total of the correct foreground pixels in the video sequence, and evaluating those relative performances with the error value (%) using a variety of video sequences. From this, we conclude why the Gaussian Mixture Model is the most used for tracking people and for all other operations of computer vision. One of the most effective and common techniques to separate the moving objects is a background subtraction, in which a model of the static scene background is subtracted from each frame of a video sequence. This technique has been applied by many researchers [1].

The task of moving objects is affected by several factors such as illumination changes, camouflage and shadows, which are the origins of misclassification in process.

In this paper, we present a calculated evaluation basing on the error % during the motion, and we normalize by using two metrics accuracy. This measurement methodology and set of metrics that we hope will help operators in industry to make more informed decisions when assessing the various video analytic solutions available on the market. Our study calculates the number of pixels in foreground object after the process of subtraction when the process of thresholding is correctly realized. Each one of the three techniques used in the paper applied a test to a foreground threshold number (for example in the Mixture of Gaussian algorithm: $0.25 < thresh < 0.75$) it is an important condition used to classify pixels in foreground/background.

The design process of mobile person detection and area entry tests through one view camera has been described as the scheme given below:

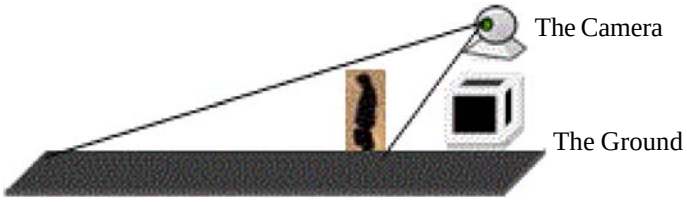

Figure 1. Tracking people with one static camera view

This work gives an efficient qualitative and quantitative evaluation of mobile people detection. It's organized into three good sections. Section 2 describes the process of separating background and foreground in an input video sequence. This evaluation

is based on computing misclassified pixels, furthermore, we give the mathematical formulation for each technique. In section 3, we present our experiment results for each background subtraction visually in real frames extracted from our video sequences, and in terms of % in order to explain the visual results. As a final point, exploiting the number of misclassified pixels, we can trigger a sign to control service when the first correct pixel is entering in our controlled area. Conclusive remarks are described in section 4.

2. Foreground separation process

In this work, we use a frames sequences which captured by a fixed camera in order to work with a stable background. We use respectively : frame difference, approximate medium and the Mixture of Gaussian for low-complexity, Medium complexity and High-complexity and project them at very different situation sequences in light, length of the sequence, position of the object to locate and different degrees of occlusion.

2.1 Frame Difference

To identify moving objects from a video sequence is a fundamental and critical task in every computer vision application. In this section we present the frame difference to subtract background from a video sequence. It is the simplest and the lowest in implementation, to separate the foreground pixels to background pixels with this approach as the difference between the current frame and an image of the scene's static background [9]:

$$| frame_i - background_i | > Th$$

The first difficulty is how to automatically obtain the background i. In order to simply and correctly estimate foreground, the estimated background is just the previous frame as explaining in the equation:

$$| frame_i - background_{i-1} | > Th$$

But it must be applied in special condition of objects speed and frame rate, and it's sensitive with high degree to the threshold parameter. The problem is that the interior pixels are interpreted as background pixels, and if an object stays for a few seconds, it is considered as black pixels (background).

2.2 Approximate Medium

In this algorithm, the previous frames of sequence are saved in buffer and calculate their median. Then, the background is subtracted from the current frame and thresholded in order to extract the foreground pixels. The median filtering method [1] works such as: if a pixel xi in the current frame is with a value larger than their corresponding background pixel xic, then the background pixel is incremented by 1. Similarly, if the current pixel is less than the background pixel, the background is decremented by 1. Practically this technique is a very good compromise because it offers performance near what you can achieve with higher complexity methods and it costs not much in computation.

2.3 Mixture of Gaussian Model for background separation

The Gaussian terms are represented by (μ_k, σ_k, ω_k) [2]. In this way, the model copes with multi-modal background distributions: The number of modes is predefined (k = 3, 4, 5).

With ω_k are the weights that are normalized and updated at every new frame, μ_k are the mean vector of the Gaussian component and k is the covariance matrix.

At every new frame, some Gaussians match the current value (all components with distance < $2.5\sigma_k k$) : for thesecomponents μ_k, σ_k are updated by the running average. The color distribution of a pixel is modeled by the *K* Gaussians components, with the term $k-th$ Gaussian presents mean vector μ_k represented by $\mu_k = < \mu_k(R), \mu_k(G), \mu_k(I) >$ in *RGI* space compared to *RGB*, *RGI* normalized color is less sensitive than *RGB*, to all small changes in illumination which is done by the shadows. With this Gaussian model, at any time, than what is known about a particular pixel $x = (x_0, y_0)$ is its history: $\{X_1, X_2, ..., X_t\} = \{I(x_0, y_0, i): 1 \leq I \leq t\}$, with I : is the image sequence.

We describe each pixel XN by the k-Gaussians [3, 4] :

$$P(X_N) = \sum_{j=0}^{k} \omega_j \eta(X_N)$$

Each pixel is modeled separately [3, 9, 10] by a mixture of *k*-Gaussians, where $k = 4$ in the paper at [5], and in [3, 4] it is assumed that $\theta_j = (i, t)2I$ and the background is updated, before the foreground pixels are detected.

Where ω_j is the weight of the kth Gaussian component, $\eta(X, \theta_j)$ is the normal distribution of k[th] component represented by [6]:

$$\eta(X_N, \mu_k, \Sigma_k) = g_{(\mu, \Sigma)}(X) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}$$

Σ is the covariance matrix, the first *B* distributions are used as a model of background of the scene by estimating *B* as [7]:

$$B = arg_b min \left(\sum_{j=1}^{b} \omega_j > T\right)$$

T is the minimum portion of the image which is expected to be background (it represents a threshold value). Then components 1... *B* are assumed to be background, if X_N does not match one of these components and the pixel is marked as foreground. All pixels in foreground are segmented into regions with using connected component labeling and detected regions are described by their centroid [10]. Then X_N is within a standard deviation α of μ (where $\alpha = 2.5$ in [4]), then we update the ith component by the following equations:

$$\omega_{i,t} = \omega_{i,t-1}$$

$$\mu_{i,t} = (1-\psi)\mu_{i,t} + \psi I_t$$

Where

$$\psi = \alpha Pr(I_t | \mu_{i,t-1}, -\Sigma_{i,t-1})$$

$$\Sigma_{i,t} = \sigma_{i,t}^2 I$$

and

$$\sigma_{i,t}^2 = (1-\psi)\sigma_{i,t-1}^2 + \psi (I_t - \mu_{i,t})^T (I_t - \mu_{i,t})$$

Where: $1/\alpha$ defines the time constant determining changes.

The background is detected as follows, all components in the mixture of Gaussian are stored into a decreasing order $\omega_{i,t} / \| \Sigma_{i,t} \|$, for all components with lowest variance and most evidence are assumed to be the background. The background subtraction is performed by the operation of marking a foreground pixel any pixel with standard deviation $\beta > 2.5$ from any of the terms of the B distributions.

An expectationâĂŞmaximization (EM) algorithm is applied to estimate the *k* and the mixture parameters of each distribution and the distribution is modeled by the Gaussians whose weights ω_k is greater than a threshold T_ω.

We use the expectation maximization algorithm (EM) [7] as an efficient iterative method to find the Maximum Likelihood (ML) estimate in the presence of missing or hidden data. In this algorithm we begin our estimation of the Gaussian mixture model by expected sufficient statistics update equation then we switch to *L*-recent window version when the first *L* samples are processed. The expected sufficient statistics update equations provide a good estimate at the beginning before all *L* samples can be collected. We use this approximate algorithm which essentially treats each new observation as a sample set of size 1 and uses standard learning rules to integrate the new data. Finally, we can sum up GMM algorithm into four steps:

- Compare the input pixels to the means μ_k.
- Update the components variables (ω, μ and σ).
- Specify which components are parts of the background.
- Extract foreground pixels

3. Eperimental results

Practically, we adopt the steps described in this diagram:

The following sections presents in two steps the results of foreground subtraction process and the output data in the room control people. We use one fixed and calibrated camera in order to have a stable background. This camera supervise one person in a room like operation room in a hospital or a person in any other sensitive application.

In the second state of this works we propose to divide the view of our camera to four parts, in order to test if the presence of the person in each region of the room.

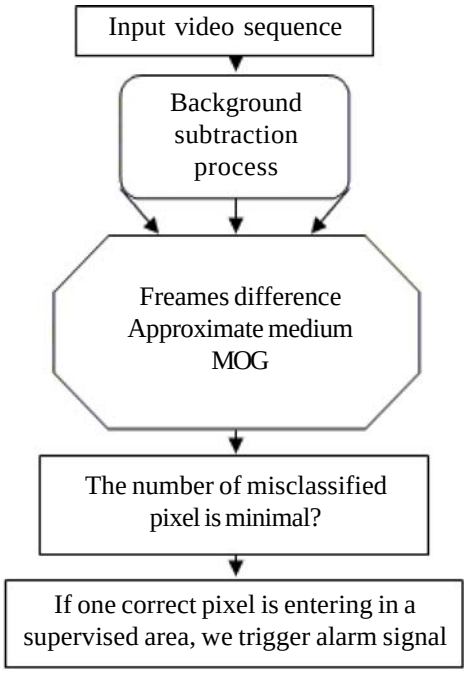

Figure 2. Overview of the algorithm

3.1 Subtraction techniques evaluation

The objective of this study is to demonstrate the efficiency of the mixture Gaussian model, in the first step of tracking people. This evaluation is modestly practiced in front of the two samples of different degrees of motion detection complexity techniques: Images difference and the approximate medium.

We use four different data sets used with different size, and video situations; all tests are given for four sequences as: In this order:

- seq 1: video PETS2006 (Ninth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance)
- seq 2: PETS with ground Truth, this video is manually created with separating the visual background from original pets
- seq 3: scene captured with webcam detected in Matlab
- seq 4: ATON from the web site address (http://cvrr.ucsd.edu/aton/shadow/ (laboratory and intelligent room)).

(a): Real frames, (b): Images different, (c): Approximate medium and (d): Mixture of Gaussian

In the results presented in the tables, we conclude that the error pixels in positives values for the Mixture of Gaussian Model are the minim. Computation value is based on the relation:

$$FPE = \frac{\sum_{n=1}^{N} P_n}{N}$$

The *FPE* metric assesses the rate of false positive values in pixels of the system used in background subtraction. Where N is the total number of frames being processed (The size of the sequence in frames) P_n is the number of false positive in the $n-th$ frame.

Figure 3. Frames result with each algorithm

FPE can take any value greater than or equal to 0. The value 0 is assumed when there are no false positives returned by the techniques when applied at the sequence, which represents the goal of any background technique. Higher values correspond to increasing false positive rates.

And we observe that the Mixture of Gaussian presents the best values (low orders) in *FPE* given in this table:

Video/Technique	Frame difference	Approximate medium	GMM
Seq 1	27.08 %	19.88 %	18.93 %
Seq 2	13.17 %	6.39 %	0.0225 %
Seq 3	6.475 %	0.728 %	0.0503 %
Seq 4	14.84 %	3.78 %	0.024 %

The *FNE* metric is obtained with the equation:

$$FNE = \frac{\sum_{n=1}^{N} F_n}{N}$$

Where *N* is the total number of frames being processed (The size of the sequence in frames).

The best value of this error is when it is equal to 0, effective and preferment background subtraction system should target to achieve a *FNE* value close to 0.

Video/Technique	Frame difference	Approximate medium	GMM
Seq 1	33,77 %	24,77 %	14,40 %
Seq 2	12,86 %	6,75 %	0,296 %
Seq 3	7,656 %	0,937 %	0,3209 %
Seq 4	14,83 %	3,55 %	0,407 %

As a result of FNE metric in Table 2, we notice that the GMM technique always takes a lower values and major all other techniques in term of minimizing value of the False Negative Error.

In this paragraph we discuss the performance of each one of the three techniques used to separate the foreground in terms of

quantitative and qualitative results applied to 4 datasets.

3.2 Qualitative and quantitative results

Figure 1 shows the visual results of each technique, (b) the results of the frames difference, (c) gives the result of the approximate medium and the (d) demonstrate the performance of the Gaussian Mixture (MoG) for background subtraction. It can be seen that the MoG performs better in terms of suppressing strong areas shadows and that the MoG gives proper objects segmented.

The quantity of error in this section is obtained by:

$$Error\,(\%) = \frac{Number\ \_\ of\_misclassified\_pixels}{Number\ \_\ of\_correct\_foreground\_pixels} \times 100\%$$

Figure 4. Real frame taken fro pets 2006

Figure 5. The four regions of background

Two metrics were used to evaluate the segmentation results. In this case we have applied the simple technique of images difference, the approximate medium and finally the Mixture of Gaussian Model in several videos. The *FPE* means that the background pixels were set as foreground while *FNE* indicates that foreground pixels were identified as background. The curves in Table 1 and 2 show the comparison between the three techniques used at different states of accuracy.

The Mixture of Gaussian Model always represents the small values of error.

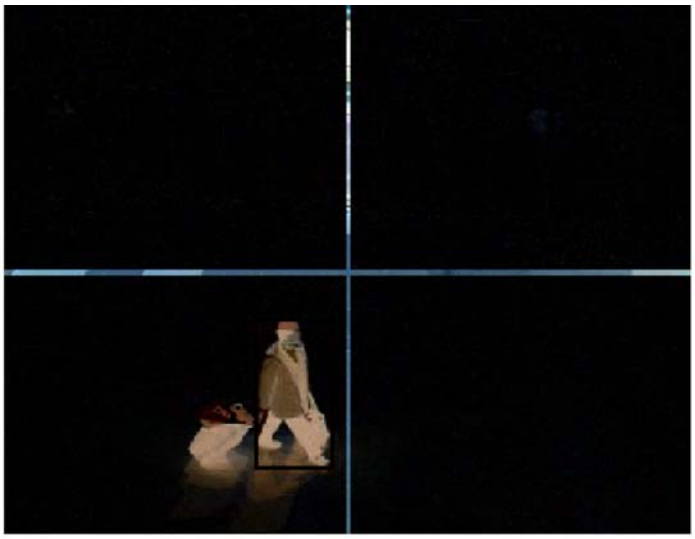

Figure 6. Real frame without background

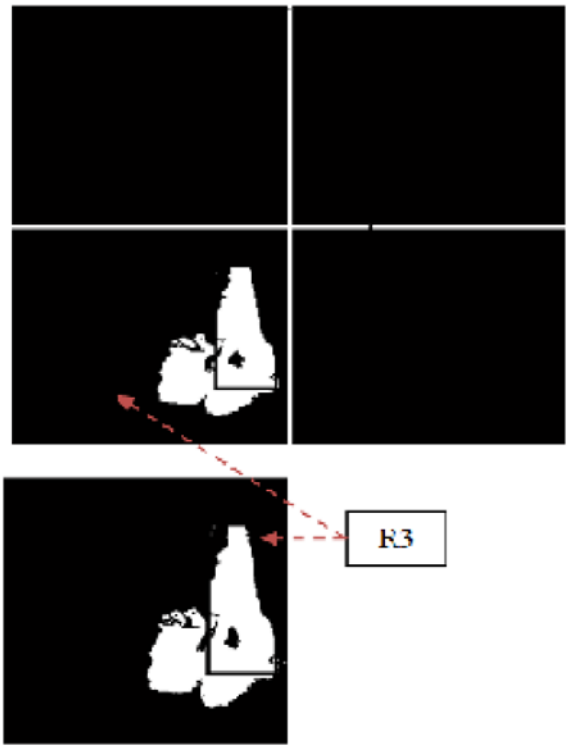

Figure 7. A person detected in region 3

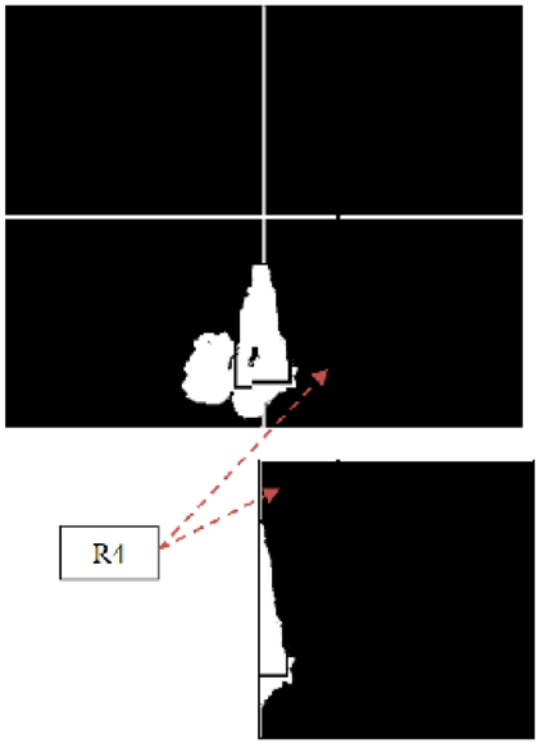

Figure 8. A Person arriving in region R4

3.3 Control of people entering a region

Using the Gaussian Mixture Model to separate the background and the foreground sections, we have to divide the background

into four regions in order to test the presence of people in video sequence viewed by a static camera which will be well calibrated to cover all the space in question:

After separating the foreground and background of the scene, we select only the third region with the person entering in the region and we test if there are one or more pixels behind the separating line. After getting the interest moving object, we extract the interest moving object by using the bounding Rectangle box which is determined by computing the maximum and the minimum values of x and y coordinates of the interest moving object.

After foreground detection matrix, if a person in control is in a sensible region, the system returns an alarm to inform the responsible unity (process under research), see figure 3 and 4.

4. Conclusion

In this paper, we have presented an evaluation technique between three methods for foreground extraction and supervising a person entering in a special area. This study demonstrates the good results in output of the statistical modeling for pixels presented by the Gaussian Mixture Model. We have concluded that the MoG is more robust, because it can handle multi-modal distributions. The MoG method has five parameters which must be tweaked (the background component weight threshold Ts, the standard D, the learning rate , the total number of Gaussian components, and the maximum number of componentsM in the background model); all of them can have a significant influence on the performance of the algorithm. After demonstrating that the MoG is the powerful technique for extraction the moving object in the scene, we apply a test of the presence of pixel in entering one region of the scene.

References

[1] Piccardi, M. (2004). Background Subtraction Techniques: a Review. *In*: Proceeding s of the IEEE International Conference on Systems, Man and Cybernetics (SMC'04), The Hague, The Netherlands, October.

[2] Thiago, Santos, T., Carlos, Morimoto, T. (2011). Multiple Camera People Detection and Tracking Using Support Integration. in ELSEVIER, *Pattern Recognition Letters*, 32, 47Ű55.

[3] Grimson, W. E. L., Stauffer, C., Romano, R., Lee, L. (1998). Using Adaptative Tracking to Classify and Monitor Activities in a Site. *Computer Vision and Pattern Recognition Santa Barbara*, California (Jun. 1998) p. 1-8.

[4] Stauffeer, C., Grimson, W. E. L. (1999). Adaptive Background Mixture Models for Real-Time Tracking. *Computer Vision and Pattern Recognition Fort Collins*, Colorado (Jun.1999) p. 246-252.

[5] Ivanov, Y., Stauffer, C., Bobick, A.,Grimson, W. E. L. (1999). Video Surveillance of Interactions. *Second IEEE Workshop on Visual Surveillance Fort Collins*, Colorado (Jun. 1999) p. 82-90.

[6] KaewTraKulPong, P., Bowden, R. (2001). An Improved Adaptive Background Mixture Model for Realtime Tracking with Shadow Detection. 2nd European Workshop on Advanced Video Based Surveillance Systems, AVBS01. Sept. VIDEO BASED SURVEILLANCE SYSTEMS: *Computer Vision and Distributed Processing*, Kluwer Academic Publishers.

[7] Chen, Y. Y.-T. et al. (2006). Efficient Hierarchical Method for Background Subtraction. Pattern Recognition, doi: 10.1016 / j.patcog. 2006.11.023.

[8] Dempster, A., Laird, N., Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39 (Series B): 1-38.

[9] Raheja, J. L., Pallab Jyoti Dutta, Sishir Kalita, Solanki Lovendra, (2012). An Insight into the Algorithms on Real-Time People Tracking and Counting System. *International Journal of Computer Applications* (0975-8887) 46 (5), May.

[10] Stein, G. P. (1999). Tracking from Multiple View Points : Self-Calibration of Space and Time. *Computer Vision and Pattern Recognition Fort Collins*, Colorado (Jun. 1999) p. 521-527.